iConference 2025

20TH INTERNATIONAL CONFERENCE

Living in an AI-gorithmic World

EDITORS:

ISAAC SSERWANGA / MADELYN ROSE SANFILIPPO / CHARLES INSKIP / ANNIKA HINZE / FRANCE BOUTHILLIER / SARA MARTÍNEZ CARDAMA / JOSEP COBARSÍ MORALES / ALAN CÉSAR BELO ANGELUCI / JULIÁN D. CORTÉS / RHEA ROWENA APOLINARIO

VIRTUAL PROGRAM MARCH 11–14

ON-SITE PROGRAM MARCH 18–22
IN BLOOMINGTON, INDIANA, USA

iSchools

LUDDY SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING
INDIANA UNIVERSITY

Preface

As we embark on the proceedings of the 20th anniversary of iConference (2025), we look back on the transformations since the iConference of 2024. This annual event is a testimony to the resilience, adaptability, and innovation of the information community. We acknowledge the pervasive technology of new algorithms, and their profound, and interconnected impact on our work and daily lives, by shaping everything from social interactions to economic activities. In this conference iSchools community has investigated impact of algorithmic processes to inspire and explore new areas of research.

The success of the iConference 2024 hinged on our commitment, shared goals, and our leveraging the power of information to enhance the well-being of individuals and communities. Since then, a new global trend has emerged that the iConference 2025 slogan encapsulates as '*living in an AI -gorithmic World.*' The iConference papers have explored the profound implications of collecting, storing, retrieving, using, and distributing information that is increasingly influenced by algorithmic processes, such as generative artificial intelligence, machine learning, data analytics, and related technologies. As we prepare the proceedings, we invite readers to immerse themselves in the evolving scholarly discourse and its contribution to the diverse shifts driven by AI, LLMs, machine learning, and predictive analytics.

The virtual iConference 2025 took place from March 11 - 14, 2025 before the physical conference on March 18 - 22, 2025 at Indiana University, Bloomington, Indiana, US. Its host was the Indiana University Bloomington in the USA.

The conference theme attracted a total of 249 submissions with 111 Full Research Papers, 133 Short Research Papers and 5 Spanish/Portuguese Research Papers.

In a double-blind review process by 451 internationally recognised experts, 103 entries were approved, including 35 Full Research Papers and 68 Short Research Papers. The approval rate was 31.53% for the Full Research Papers and 51,13% for the Short Research Papers. Additional submissions were selected for the workshops and panels, the doctoral colloquium, the early career colloquium, the student symposium, posters, and the Spanish-Portuguese and Chinese language paper sessions.

The Full, Short and Spanish /Portuguese Research papers will be published for the first time open access in Information Research. These proceedings are organised into eighteen categories, reflecting the diversity of the information research areas: '*behavioural research*', '*ai and machine learning*', '*information science and data science*', '*information and digital literacies*', '*cultural heritage, archives and museums*', '*information science education*', '*information governance and ethics*', '*open access and open science*', '*digital humanities*', '*bibliometrics and metadata*', '*disinformation and misinformation*', '*social media and digital networks*', '*libraries and users*', '*health informatics*', '*human-ai collaboration*', '*information retrieval*', '*community informatics*', and '*scholarly communication*'.

We are deeply thankful to the reviewers for their expertise and invaluable contributions to the review process, and to the track chairs for their dedication and expert knowledge. We wish to extend our gratitude to the chairs and volume editors: Full Research Papers chairs Madelyn Sanfilippo from University of Illinois Urbana-Champaign United States, Charles Inskip from University College London United Kingdom, and Annika Hinze from University of Waikato New Zealand; Short Research Papers chairs Rhea Rowena Ubana-Apolinario from University of the Philippines Philippines, and France Bouthillier from McGill University Canada; Spanish and Portuguese Research Papers chairs Sara Martínez Cardama from Universidad Carlos III de Madrid Spain, Josep Cobarsí Morales from Universitat Oberta de Catalunya Spain, Alan César

Belo Angeluci from Universidade de São Paulo Brazil, and Julián D. Cortés from Universidad del Rosario Colombia.

The iConference successfully reflected the global representation of iSchools and fostered the synergy between research and teaching in information and complementary areas of AI.

February 2025

Isaac Sserwanga
Madelyn Sanfilippo
Charles Inskip
Annika Hinze
Rhea Rowena Ubana-Apolinario
France Bouthillier
Sara Martínez Cardama
Josep Cobarsí Morales
Alan César Belo Angeluci
Julián D. Cortés

Organisation

Organizer

Indiana University Bloomington United States

Conference chairs

Howard Rosenbaum.	Indiana University Bloomington United States
Carol Choksy.	Indiana University Bloomington United States
Devan R Donaldson.	Indiana University Bloomington United States

Proceedings chair

Isaac Sserwanga	iSchools Organisation

Full research paper chairs

Madelyn Sanfilippo	University of Illinois Urbana-Champaign United States
Charles Inskip	University College London United Kingdom
Annika Hinze	University of Waikato New Zealand

Short research paper chairs

Rhea Rowena Ubana-Apolinario	University of the Philippines Philippines
France Bouthillier	McGill University Canada

Poster chairs

Eric M. Meyers	The University of British Columbia Canada
Laura Sbaffi	University of Sheffield United Kingdom
Ping Wang	Central China Normal University China

Spanish - Portuguese papers chairs

Sara Martínez Cardama	Universidad Carlos III de Madrid, Spain
Josep Cobarsí Morales	Universitat Oberta de Catalunya, Spain
Alan César Belo Angeluci	Universidade de São Paulo, Brazil
Julián D. Cortés	Universidad del Rosario Colombia

Chinese paper chairs

Rong Tang.	Simmons University United States
Tina Du.	Charles Sturt University Australia
Pengyi Zhang.	Peking University China

Workshops and panel chairs

Bei Yu	Syracuse University United States

Rahmi	Universitas Indonesia Indonesia
Helena Lipková	Charles University Prague Czechia

Student symposium chairs

Anika Meyer.	University of Pretoria. South Africa
Hiroyoshi Ito.	University of Tsukuba Japan
Matthew Francisco.	Indiana University Bloomington United States

Early career colloquium chair

Udo Kruschwitz	University of Regensburg Germany
Dion Hoe-Lian Goh	Nanyang Technological University.Singapore
Kristin Eschenfelder.	University of Wisconsin-Madison United States

Doctoral colloquium chair

Güleda Doğan	Hacettepe University Turkey
Khalid Hossain.	Monash University Australia
Heather Moulaison-Sandy.	University of Missouri United States

Conference coordinators

Noriko Hara.	Indiana University Bloomington United States
Gillian Dunn.	Indiana University Bloomington United States
Slava Sterzer	iSchools Organization
Vera Hillebrand	iSchools Organization
Ulrike Liebner	iSchools Organization

Reviewers Full and Short Papers iConference 2025 (429)

Jacob	Abbott	Hui	Chambers
Janai	Adams	Michelle	Chao
Noa	Aharony	Minghong	Chen
Eiman	Ahmed	Xiang	Chen
Shameem	Ahmed	Xiaoyu	Chen
Isola	Ajiferuke	Yi-Yun	Chen
Nicole D.	Alemanne	Shih-Yi	Chen
Daniel Gelaw	Alemneh	Gi Woong	Chen
Reham	Alhuraiti	Jeongbae	Chen
Obianuju	Aliche	Wonchan	Chen
Lilach	Alon	Yujin	Cheng
Iman	Alshathri	Yunseon	Chien
SYDNEY	An	Gobinda	Choi
Misita	Ann Subin	EunKyung	Choi
Rhea Rowena Ubana	ANUYAH	Malcolm John	Choi
MAYUKH	Anwar	Raphael	Choi
Jessica	Apolinario	Mónica	Choi
Zoe	BAGCHI	Andrea	Chowdhury
Edward A	Ball	Isabella L.	Chung
Elmi	Barfield	Laura	Clark
Kayla M	Bartliff	Chris	Cobe
Christine L.	Benoit	Andrew	Colón-Aguirre
Theo JD	Bester	Sally Jo	Copeland
Guillaume	biancapeterson	Diana	Corieri
Paulina	Booth	Elijah John	Costa
Jo Ann M.	Borgman	Gabriel	Costello
Yi	Bothma	Nilou	Coward
Sarah A.	Boutard	Alison	Cox
Julia	Bressel	FNU	Cunningham
Mimi	Brooks	Jun	Daly
Yao	Bu	Shengli	Dar Juan
Aldrin Joseph	Buchanan	Junhua	David
Yu	Bullard	Ly	Davoudi
Zhe	Byun	Güleda	Day
Daniel	Cai	Rhys	Dedema
Sunandan	Campos	Yunfei	Deng
Tamy	Cao	Kedma	Deng
Tiffany	Cao	Elizabeth	Ding
Haihua	Carter	RICARDO	Dinh
Hsinliang	Chakraborty	Avshalom	Doğan

Tao	Dozal
Zhenjia	Dreeszen Bowman
Katrina Simone	Du
Rachel	Duarte
Fred	Duarte dos Santos
Ina	Eikey
Rebecca D.	EITO-BRUN
Darin	Elmalech
Viviane	Enam
Hengyi	Enwald
Yaming	Fan
Henry	Fan
Maria	Fenlon
Chunmei	Fleming-May
Gao	Fonseca
John	Fourie
Sanjana	Frank
Rich	Freeburg
Yegin	Frings-Hessami
Diane	Fu
Dion H.	Fu
Anne	Gabb
Rob	Gäde
Elke	Gadi
Jessica Helen	Gan
Melissa	Gao
Qian	Gathegi
Neslihan	Gautam
Ayse	Gazan
Lala	Genc
Ariel	Gill
Preben	Goh
Yue	Goulding
Noriko	Grace
Jean	Greifeneder
Bruce	Grimmer
Morgan	Gross
Stefanie	Gryk
Daqing	Gumusel
Jiangen	Guo
Susan C.	Gurol
Alison	Gursoy
Darra Lynn	Hajibayova
Chris	Hammond
Md Khalid	Han
Jianhua	Hansen
Jingrui	Hao
Wenjun	Hara
Amanda	Hardy
TSUNG-MING	Hartpence
Qibiao	Harvey
Xinhui	Havelka
Yuerong	He
Zhan	He
Zhiwei	He
Ching Yin	Hellmer
Ruhua	Herring
Ying	Hicks
Yuting	Hitt
Shezin	Hofman
Isto	Holstrom
Jiwoo	Hossain
Tomoya	Hou
Charles	Hou
Yvette	Hou
Emi	Hovious
Jonathan	HSIAO
Corey	Hu
Fariha	Hu
Piyapat	Hu
Robert	Hu
Ellie (Eunmi)	Hu
Zhuoren	Huang
Jenny	Huang
Erik	Huang
Michael	Huang
Nicolas	Huang
Jaap	Huang
Jumana Yousef	Hussain
Anajoyce Samuel	Huvila
Ijay	Hwang
Mat	Igarashi
Rebecca	Inskip
Heikki	Iribe Ramirez
Mahmood	Ishita

Mohammad Golam Isip
Jiro Ito
Halil Jackson
Jeonghyun Jaigirdar
Vanessa Lynn Jarusawat
Kyungwon Jäschke
Masanori Jeong
Kushwanth Jiang
Maja Jiang
Ravi Johannisson
Mucahid Johnston
Jessica M. Jones
Chengyi Jullien
Anna R. Kamps
Deborah Kass
Hyun Seung Katabalwa
Kijung Kaz-Onyeakazi
Myeong Kelly
Tae Hee Kelly
Honglei Keskustalo
Lei Khosrowjerdi
Longfei Kibria
Wenqi Kikkawa
Xin Kilicoglu
xinyue Kim
Ying Kitzie
Yingya Koh
Jingwen Koizumi
Zhixuan Koya
Lizhen Krtalic
Shaobo Krueger
Louise Kuber
Jenny Kutlu
Zack Lapp
Onkar Rupesh Layne-Worthey
Chang Le, Chengyi
Jieli Leach
Jing Lee
Ping Lee
Xiaohui Lee
Yaqi Lee
Annemaree Lee
Kun Li
António Li
Haakon Li
Shutian Li
Emily Li
Pattiya li
Alaine Li
Keith Li
Kathryn Lian
Masaki Lian
Miriam Liang
Matthew Liang
Regina Dwi Shalsa Limberg
Suvodeep Lindberg
Pamela Ann Lischer-Katz
David Litake
Melissa Liu
Shuyuan Liu
Anika Liu
Eric Liu
Shawne D. Liu
Stacie Liu
A.J. Lloyd
J Elizabeth Lu
Stasa Lucas Soares
Yue Lund
Yosuke Luo
Robert Ma
Atsuyuki Ma
Heather Maemura
Cameron Mahapasuthanon
Anirudh Martaus
Sarah Marzullo
David M. Masten
Rebecca Matsubara
Dian Matteson
Kathleen Lourdes Mayernik
Melissa G. Mayzana
Kyong Eun Mazumdar
Benedict Salazar McKinney
Gillian McMenemy
Peter Meierhofer
Felipe Melo
Giulia Metcalfe

Shreya
Weimei
Velian
Hyoungjoo
Min Sook
Natalie
Lei
Rebecca
Sean
Olivia
Bobby
Ola
Leonor Gaspar
Nayana
Alex
Gila
Xin
Jian
Rahmi
Arcot
Jelena
Diane
Susan
Christopher B
Benhur
Alexandria Jane
Ming
Cristian
Howard
Vassilis
Mona
Özhan
SONGLAK
Rachel
Jomara
Madelyn Rose
Kirsten
Michael
Subhasree
Hugh
Sarika
Moushumi
Ryan

Meyer
Meyers
Miksa
Milburn
Million
Mills
Milojevic
Ming
Miyata
Mohamad
Montoya
Morishima
Moulaison-Sandy
Moy
Murillo
Nagraj
Nakash
Nesset
Nguyễn
Nichols
Noone
Novita Fitriani
Obille
Ocepek
Oh
Olgado
Oliver
Organisciak
Ortega
Osti
Paithankar
Pan
Pandeliev
Park
Park
Parker
Pei
Penrose
Pessin
Pestana
Phuritsabam
Pilerot
Pinto

Qinggong
Erica
Henrique Diogo
Stephen C.
Clay
Beth
Gretchen Renee
Rebecca
Owen
Besiki
Meicen
RAN
Shaodan
zhuo
Allan
Sue Yeon
Masao
Aiko
Anna Maria
Jian
Rong
Meriem
Paula
Kuruppu Arachchige
Chunhua
Tien-I
Zhifang
Rachel
Michael
Diego
Martie (Martha Johanna)
Brenda
Merce
Di
Fan
Jieshu
Jieyu
Pianran
Xiangnyu
Xinyue
Xuyan
Yanyan

PM
Poole
Prebor
Qian
Qin
Rahmi
Rajasekar
Rajić
Rasmussen McAdie
Rathbun Grubb
Rauch
Ravuri
Rayburn
Ren
Roman Palacios
Rosenbaum
Routsis
Ryan
Sachdev
Saglik
Sakulwichitsintu
Salzano
Sandbulte
Sanfilippo
Sbaffi
Schlebbe
Scully-Blaker
Seadle
Sengupta
Shanahan
Sharma
Sharmin
Shaw
Shi
Shusas
Silva
Simte
Slota
Spinuzzi
St. Jean
Stahlman
Stallworth

Yuhua	Stewart-Robertson
Zhongyi	Stvilia
zixin	Sun
Karen M.	SUN
Dietmar	Sun
Dawei	Sun
I-Chin	Suzuki
Shuheng	Svarre
Peng	Sylvester
Haoyu	Syn
Iris	Takaku
HUIMIN	Takazawa
Lei	Tammaro
Lifang	Tang
Shenmeng	Tang
Yanrun	Tebourbi
Jian	Telo
Hui	Thusitha Chamara
Lijuan	Tian
Naiyi	Tsai
Ruixian	Tsai
shengnan	TU
Junling	Tunis
Ayoung	Twidale
JungWon	Vallejo-Huanga
Sarah	VanDeventer
Fei	Van Wyk
Fengchang	Vazquez
Samaneh	Wang
Xianjin	Wang
Yujia	Wang
Chenwei	Wang
Li	Wang
Pengyi	Wang
Tongyang	Wang
Xiaoqian	Wang
Yishan	Wang
Yuhao	Wang
Liang	Wang
Yuxiang (Chris)	Wang
Heng	Wang
Kyrie Zhixuan	WANG
Lihong	Wang
Qinghua	Wang
Zixin	Wang

Reviewers Spanish and Portuguese papers (23)

Lillian	Alvares
Vania Mara	Alves Lima
Rogério	Aparecido Sa Ramalho
Marielle	Barros de Moraes
Paula Carina	de Araújo
Juan Carlos	Fernández Molina
Francisco Javier	García Marco
Montserrat	Garcia-Alsina
Liliana	Giusti Serra
Carlos	Guardado da Silva
Eduardo	Juarez-Valero
Angelica	Marques
Carla Maria	Martellote Viola
Sandra Gisela	Martin
Lisandra	Otero Borges
Pablo	Parra Valero
Fernanda	Ribeiro
Blanca	Rodriguez Bravo
Ricardo	SantAna
Brenda	Siso-Calvo
Paula	Telo
Ana Lúcia	Terra
Maria Cristina	Vieira de Freitas

Table of Content

Behavioural Research

The promoting effect of internet use on the physical health of Chinese older adults: an empirical study based on the difference-in-differences model 1

Haoyuan Sun, Zhenkang Fu, Zhengtong Pu, Qinghua Zhu

Navigating end-of-life decision making in a small world: a thematic analysis of Uchinanchu American older adult's information behavior.....11

Kristina Shiroma

'You can't imagine how hard it is': hardship streamers in live crowdfunding.....23

Yueru Yan, Pnina Fichman

Moral motivational duality and fake news reporting behavior: a polynomial regression with response surface analysis.....38

Xiao-Liang Shen, Lin-Yao Liu, Yang-Jun Li

Drivers of metaverse adoption among children and youth: a UTATU2 perspective 54

Liuyu Huang, Dion Hoe-Lian Goh, Stella Xin Yin

Ética, Política e Informação: uma reflexão conceitual sobre sociedade e comportamento.....63

Vitor Serejo Ferreira Batista

AI and Machine Learning

Evaluating techniques of artificial intelligence for social robots.....74

Jessica K. Barfield

Taking Disagreements into Consideration: Human Annotation Variability in Privacy Policy Analysis, ...81

Tian Wang, Yuanye Ma, Catherine Blake, Masooda Bashir, Hsin-Yuan Wang

A more advanced group polarization measurement approach based on LLM-based agents and graphs 93

Zixin Liu, Ji Zhang, Yiran Ding

The role of ontologies in machine learning: a case study of gene ontology.....108

Qiaoyi Liu, Jian Qin

Finding Pareto trade-offs in fair and accurate detection of toxic speech, ...123

Soumyajit Gupta, Venelin Kovatchev, Anubrata Das, Maria De-Arteaga, Matthew Lease

Safety anchors and deflected desires: generative AI and the production of sexualities.....142

Daniel Carter

Empowering customer service with generative AI: enhancing agent performance while navigating challenges 150

Charles Costa, Souvick Ghosh

AI lifecycle from a data-driven perspective: a systematic review 159

Di Wang, Ruiyang Chen, Chuanni Li, Shanshan Gu

Information science and data science

Gallery walk as research method in information science.....171

Rebecca Frank, Stephanie Krueger

Discourses of fear around AI and their implications for library and information science 180

Sarah Appedu, Yigang Qin

Transforming bridges into smart infrastructure: a data-driven approach to monitoring bats and promoting sustainable coexistence 189

Jacob Koch, Jess Kropczynski, Janette Perez-Jimenez, Joseph S. Johnson

Improving scholarship accessibility with reinforcement learning 203

Haining Wang, Jason Clark, Hannah McKelvey, Leila Sterman, Gao Zheng, Zuoyu Tian, Xiaozhong Liu

How do data authors perform in data-intensive research activities? Evidence from author contribution statement in data papers 219

Yang Heng, Yu Yonglin, Liu Fenghong

A benchmark for evaluating crisis information generation capabilities in LLMs.....240

Ruilian Han, Lu An, Wei Zhou, Gang Li

Information and digital literacies

Design of audio ads to prebunk misinformation and promote civil discourse.....249

Diana Daly, Kainan Jarrette

Beyond the hype: longitudinal trends in virtual reality perceptions beyond the COVID-19 pandemic.....260

Christopher Ball, Kuo-Ting Huang, Jess Francis-Levin

Engaging with information beyond vision: hands-on approaches to computational thinking for blind and visually impaired learners 280

Kyungwon Koh, JooYoung Seo, Si Chen, Eugene Malcolm Cox

Cultural heritage, archives, and museums

Does Google dream of electric memes? From human to computational culture..... 287

Alexander O. Smith, Una Joh, Jeff Hemsley

Teaching computational archival science: context, pedagogy, and future directions.....301

Victoria L. Lemieux, Richard Marciano

Gendered recordkeeping practices in marginalised communities in Bangladesh 319

Viviane Frings-Hessami

The affective dimension of archival work: understanding the thoughts and feelings of archivists who documented the HIV/AIDS epidemic 330

Allan A. Martell, Travis L. Wagner, Shannon Oltmann

An ensemble framework for sentiment-embedded event evolution in diaspora oral archives....338

Jing Zhou

Innovative practice of archival data development workflow in the AGI era: a case study of scientist archives project 349

Yaming Fu, Jie Song, Xinran Zhang, Jingyun Bi

Data feminism in museum collections: mapping principles to practices.....361

Alexandria Rayburn

A semantic framework for digital archives of ancient Chinese villages aimed at humanistic discovery: a case study of four villages in Henan.....373

Tianjiao Qi, Yiduo Wang, Meiman Wong, Tsamchu Basang, Linglu Li

A cultural heritage data curation framework for knowledge discovery.....398

Ruijie He, Xiaoguang Wang

Ideological presentations in official promotion of intangible cultural heritage (ICH) on short-form video platforms: a multimodal content analysis 413

Ruoxi Liu, Haoran Qiu

Information science education

Professional identity scale for information professionals: an instrument development 425

Fatih Oguz

Stewardship of digital language archives: training development and testing through collaboration of information scientists, linguists, and communities 435

Oksana L. Zavalina, Alexandra O'Neil, Shobhana L. Chelliah

The inevitability of AI: a study of undergraduate students' perceptions of AI tools in their future Careers 443

Mónica Colón-Aguirre, Kawanna Bright

The smart archive management practicing pipeline: a virtual online learning platform for AI literacy development in archive science 450

Lanxi Dong, Shuning Tang, Yuan Cheng, Ping Wang

Information governance and ethics

'Just because we can, does it mean we should?' The integrated data infrastructure in Aotearoa, and its implications for Māori 467

Daphne Boey, Jennifer Campbell-Meier, Spencer Lilley, Janet Toland

The impact of privacy authorization framing in mobile apps on privacy disclosure: a regulatory focus approach 476

Zhuo Sun, Dandan Zhang, Xiaoning Gao, Yan Jin

Does artificial intelligence harm labour? Investigating the limitations of incident trackers as evidence for policymaking 486

Theodore Dreyfus Ledford

Cultivating university data culture in the age of artificial intelligence: a conceptual framework and critical reflections 500

Yunjie Tang

'Sora is incredible and scary': public perceptions and governance challenges of text-to-video generative AI models.....508

Kyrie Zhixuan Zhou, Abhinav Choudhry, Ece Gumusel, Madelyn Rose Sanfilippo

Investigating privacy risks in open government data: an exploratory case study.....523

Daniel Carter, Caroline Stratton

Using the information inequity framework to study GenAI equity: analysis of educational perspectives. 533

Sarah Zipf, Chuhao Wu, Tiffany Petricini

Governing knowledge commons in information science 548

Ebubechukwu E. Uba, Madelyn Rose Sanfilippo

Steering the AI world: an exploratory comparison of ai acts in the EU and Canada 556

Ruiyi Zhu, Tien-I Tsai

Open access and open science

Open access: Māori perspectives on publishing.....565

Spencer Lilley

BSLISE directory: innovating global mapping and organization of LIS schools 573

Tiago Emmanuel Nunes Braga, Josir Cardoso Gomes, Nathaly Cristine Leite Rocha, Fatih Oguz, Chris Cunningham

Digital humanities

Towards nuanced and critical data analysis: an exploratory investigation into complexities of Goodreads reviews for children's books 582

Yuerong Hu and Chen Ling

Conceptual approaches to information-as-potentiality 596

Alexandra Chassanoff, Annie T. Chen

An analysis of poet demographic and thematic diversity in a poetry collection for inclusive AI..610

Kahyun Choi, Gyuri Kang

Towards memetic legitimation of knowledge: Memes and cultural heritage 618

Alexander O. Smith, Calvin Cousin, Una Joh, Christy Khoury, Yiran Duan, Jeff Hemsley

Innovation hesitancy: Exploring reluctance to digital innovation in the Swedish cultural sector.....627

Hanna Carlsson, Ahmad Kamal, Charlie Järpvall

How does digital humanities research talk about AI? A bibliometric analysis 635

Wenyi Shang, Rongqian Ma, Heather Moulaison-Sandy

The dual impact of digital humanities: evaluating the role of DH approaches in historical and cultural resource development 646

Lu Sijia, Long Jiaqing, Xu Yongjun

Bibliometrics and metadata

Systematically modeling and extracting bibliographic metadata of power grid standard documents with LLMs 654

Guowei Chen, Wei Xie, Yanan Liu, Xiaoqun Yuan, Liang Zhao

Demystifying perception and decision-making in peer review: a semantic perspective.....666

Yujie Zhang, Weikang Yuan, Zhuoren Jiang

How is AIGC shaping the world: an analysis of bibliometrics. 679

Ying Guo, Fengchang Yu, Jianfei Lai, Xiaoqun Yuan

Review articles as windows into knowledge accumulation: the case of AI research 690

Siyuan Peng, Lei Hu, Jingrui Hou, Youqing Xia, Ping Wang

Exploring the patterns and influencing factors of emerging technology impact: a case study of digital medical technology 699

Man Jiang, Siluo Yang, Shuang Tao

Leveraging scientific knowledge for technological inventions: insights from knowledge recombination theory.....712

Guiyan Ou, Haodong Chen, Kaili Wang, Jiang Wu

Are we there yet? Evaluation of AI-generated metadata for online information resources.....732

Vyacheslav Zavalin, Oksana L. Zavalina

Disinformation and misinformation

Scientists, but deny science? climate change sceptics network on YouTube led by scientists....741

Qiaoyi Liu, Yuheun Kim, Jeff Hemsley

Uncovering strategies for identifying deepfakes. 752

Celene Neo, Dion Hoe-Lian Goh, Chun Wan Ying Rachel, Chei Sian Lee

Decoding How the Older Adults Identify Misinformation: An Analysis of Internal Mechanism....761

Wu Zhaotong, Wang Yepei, Yan Hui

Social media and digital networks

Examining privacy concerns and conversations before, during, and after the onset of the COVID-19 pandemic: an analysis of r/privacy.....770

Erica Shusas, Shruti Sannon, Katie Teitelbaum, Patrick Skeba, Eric P. S. Baumer, Andrea Forte

'Everyone has their reasons for curating the data they have decided to keep': a thematic analysis of data hoarding as digital curation practice 789

Emily Maemura, Travis L. Wagner

Graphicon evolution on the Chinese platform Bilibili: a historically informed socio-technical approach 798

Yiqiong Zhang, Susan C. Herring, Yanmin Wu, Rongle Tan

Analyzing the language of rejection: a study of user flagging responses to hate speech on Reddit.....815

Sharon Lisseth Perez, Xiaoying Song, Lingzi Hong

Comunicação científica e mediação da informação em mídia social: relato de experiência dos eventos organizados pela Comissão Diálogos de Pesquisa do PPGCI/UFF.....824

Camilla Castro de Almeida, Juliana Maia Mendes, Vinicius Ribeiro Soares dos Santos, Elisabete Gonçalves de Souza, Lucia Maria Velloso de Oliveira

Libraries and users

Mapping GLAMs: creating a national dataset of GLAMs to develop a categorical climate-change risk assessment scale. 836

Edward Benoit, III, Jill Trepanier, Jennifer Vanos, Emily Fisher, Kaitlyn Bailey, Annie Waddell, Haley Moore, Mandy Hatman, Mary Sidwell, Symonne Russell, Virginia Seger, Paige Boutte, Amanda Latta, Zoe Mohammad, Kyriel Felton, Erin Deliman, Breanna Benson-Pearce, Wendy Johnson, Allyson Russell, Baillie Pretzer, Christopher Reeder, Melissa McConnell, Lisa Dahlke, Kaitlynn Melear, Lillian Bodi, Savannah T. Lyle, Zach Lannes, Gwen L. Wells, Benjamin A. Teincuff, Jason M. Straight, Tiffany Rockwell, Shane T. Manthei, Jennifer L. Benner, Jane Fiegel, Amanda Lima, Elizabeth Rininger, Caroline Melinger, Deborah Metz-Andrews, Meryl Roepke, Karen Isaac, Mallory Collins

Promoting AI literacy in U.S. academic libraries: analysing LibGuides from ARL and Oberlin group libraries using the EDUCAUSE AI literacy framework.....847

Ko Chun Ru, Rong Tang

Readers' advisory guide development for Webtoon 866

Hyerim Cho, Denice Adkins, Hillary Gould, Katy Bond, Ky Flood

Exploring psychological mechanisms of library anxiety from the perspective of achievement motive and goal orientation 875

Masaki Takeda, Satoru Suto

Towards a sustainable automated library operation; the imperatives of digital skills and related factors of digital skills development of university library professionals in Nigeria 884

Gbenga Adetunla, Gobinda Chowdhury

Digital skills development of university library professionals in Nigeria: testing the UTAUT Model.....897

Gbenga Adetunla, Gobinda Chowdhury

Aprendizaje-Servicio como instrumento para la accesibilidad y la inclusión: resultado de una propuesta colaborativa entre universidades y bibliotecas 910

Elena Loreto Olmedo-Pagés, Rosario Arquero-Avilés

Health informatics

Leveraging Social Circles and Algorithmic Processes in Digital Mental Health Tools for College Students' Stress Management.....919

Farhanuddin Fazaluddin Kazi, Jomara Sandbulte

The role of paralinguistic digital affordances (PDAs) in online mental health communities (OMHCs) in Saudi Arabia 941

Sara Alrashed, Peter A Bath, Kushwanth Koya

Enhancing vaccine communication in social Q&A: identifying readily applicable factors for answer acceptance on medical sciences stack exchange 960

Hengyi Fu

Barriers and facilitators to documenting social needs in electronic health records: a pilot survey 970

Kuo-Ting Huang, Ravi Patel, Sharon Connor, Elaine, Kong, Xuan Nie, Jocelyn Schmersal

The challenges of cross-organizational patient data sharing in medical consortium: a study within the context of two-way referral systems 982

Lihong Zhou, Hanyu Miao, Ruiyang Tu, Jiangfeng Hu

Human-AI collaboration

Collaborative human-AI risk annotation: co-annotating online incivility with CHAIRA.....992

Jinkyung Katie Park, Rahul Dev Ellezhuthil, Pamela Wisniewski, Vivek Singh

Deskilling and upskilling with AI systems 1009

Kevin Crowston, Francesco Bolici

Inconsistency-driven approach for human-in-the-loop entity matching 1024

Hiroyoshi Ito, Takahiro Koizumi, Ryuji Yoshimoto, Yukihiro Fukushima, Takashi Harada, Atsuyuki Morishima

The effects of anthropomorphic framing on senior news consumers' attitudes towards health AI: a mediation of psychological distance 1039

Yanrun Xu and Tingting Jiang

Understanding entangled human-technology-world relations: use of intelligent voice assistants by older adults.....1049

Alisha Pradhan, Shaan Chopra, Pooja Upadhyay, Robin Brewer, Amanda Lazar

Research on the influencing mechanism of blind or visually impaired persons' evaluation on generative AI in visual tasks.....1064

Huitong Chen, Yuting Pan, and Hui Yan

Will potential threats persist? Attitudes and behavioral responses of creative professionals toward GenAI 1073

Mingxia Jia, Yuxiang (Chris) Zhao, Xiaoyu Zhang

Taming TikTok: how BIPOC individuals perceive and interact with algorithmically generated content 1084

Jiarun Dai, Naila Hajiyeva, Sehba Wan, Kayla Booth

Examining generation Z's use of generative AI from an affordance-based approach 1095

Chei Sian Lee, Li En Tan, Dion Hoe-Lian Goh

Information retrieval

On the robustness of cover version identification models: a study using cover versions from YouTube 1103

Simon Hachmeier, Robert Jäschke

Exploring the role of large language model in collaborative travel planning task 1123

Hikaru Kumamoto, Hideo Joho

Information places: myth, affect, and gaps in information retrieval modeling 1131

Nathan R. Johnson

Detecting text-level intellectual influence with knowledge graph embeddings 1142

Lucian Li, Eryclis Silva

Community informatics

Information access via voice commands on YouTube: empirical evidence on the consequences for a marginalised community in Bangladesh.....1153

Joy Bhowmik, Viviane Frings-Hessami, Gillian Oliver, Md. Khalid Hossain

Beyond language archives: proposing the archival community informatics framework as an interdisciplinary link to revitalization lexicography.....1169

Merrion Dale Frederick, Ana Roeschley, Oksana L. Zavalina

From pixels to perspectives: exploring perceptions of representation in character design.....1180

Nani Le Tang, Erin Eungyo Moon, Jose De Jesus Garrido Iniguez, Daniel Raul Mena Ramirez, Daniel L Gardner

State literacy: lessons from social struggle in the Colombian Amazon 1202

Claudia Grisales-Bohórquez

Analysis of the integration of e-commerce and rural lives in China.....1209

Guangchun Zheng, Sharifa Sultana, Kate Williams

Narratives for healing: an exploration of interactions within an online 355community on navigating bullying
trauma....1217

Subhasree Sengupta, Yin Yang

Scholarly communication

Are data papers cited as research data? Preliminary analysis on interdisciplinary data paper citations 1225

Kai Li, Pao-Pei Huang, Wei Jeng

'A Mulligan's Stew': educational preparation for today's academic library liaisons in the humanities and social sciences 1234

Rachel Fleming-May, Wade Bishop, Caroline Villarreal

AI literature review systems: an analysis of performance, affordances, and outputs for a complex topic in the social sciences 1244

Heather Moulaison-Sandy, Wilson Castaño-Muñoz, Laura Ridenour, Denice Adkins

Get the root of research misconduct: a systematic review of its cultural causes....1253

Wei Feng, Lihong Zhou, Junmin Xiao

Researchers' reflections on values, mechanisms, and dynamics of knowledge and information sharing in international academic projects 1276

Dorota Rak, Mahmood Khosrowjerdi, Livhuwani Muthelo, Bhaskar Mukherjee

Information Research - Vol. 30 No. iConf (2025)

The promoting effect of internet use on the physical health of Chinese older adults: an empirical study based on the difference-in-differences model

Haoyuan Sun, Zhenkang Fu, Zhengtong Pu, and Qinghua Zhu
DOI: https://doi.org/10.47989/ir30iConf46894

Abstract

Introduction. The widespread adoption of information technology (IT) has had a profound impact on economic and social development. With China confronting an increasingly serious aging population issue, it becomes critical to examine the effects of Internet use on the physical health of older adults.

Method. This study seeks to investigate the impact of Internet use on older adults' physical health by employing a pooled cross-sectional difference-in-differences (DID) model, drawing on data from the China General Social Survey (CGSS) covering the years 2012 to 2021.

Conclusion. The results demonstrate that Internet use significantly enhances the physical health of older adults, with these findings proving robust. Mechanism tests further reveal that Internet use contributes to improved physical health by increasing the frequency of physical exercise. Additionally, heterogeneity analysis indicates that the positive effects of Internet use are more pronounced among older adults in economically developed regions and those with poor self-perceived health.

Introduction

With the swift progression of China's economy and ongoing social changes, the issue of population aging has become increasingly significant. Data from the National Bureau of Statistics (2024) shows that the number of individuals aged 60 and over in China has reached 297 million, making it one of the fastest-aging nations in the world. As people grow older, the overall health of the older adults tends to deteriorate. It is estimated that more than 190 million older adults in China are affected by chronic illnesses, with over 40 million experiencing full or partial disability (Ma, 2022). Addressing the health concerns of this large older adult population has therefore become a pressing social challenge. However, in the short term, current healthcare resources are inadequate to meet the substantial medical needs of this demographic. The ***Healthy China Plan 2030***, a key national health strategy, has consistently highlighted the importance of information technology in providing support and guidance. As a result, the Internet has emerged as a promising approach to help address the physical health challenges faced by older adults.

The Internet has not only profoundly transformed the lifestyles of older adults but has also had a far-reaching and significant impact on their physical health. While some people worry that using the Internet may have a negative impact on the health of the older adults by making their living habits worse (Gilleard et al., 2007), numerous studies have refuted this notion. Research demonstrates that Internet use does not reduce the amount of time older adults dedicate to exercise; on the contrary, it significantly increases their likelihood and frequency of engaging in physical activities(Guo et al., 2022; Sasaki et al., 2022). Moreover, older adults can access health-related information online, which enhances their health awareness and self-management abilities (Ang et al., 2021; Han & Zhao, 2021). From a methodological perspective, most current studies adopt an internal view of information behaviour, utilizing models such as ordinary least squares (OLS) regression to quantify the promoting effects of Internet use on the physical health of older adults. However, this approach has certain limitations: first, it tends to overlook external factors such as the implementation of national macro policies, the widespread availability of Internet infrastructure, and changes in the channels through which the Chinese population accesses information. Second, issues related to endogeneity between variables may compromise the accuracy and reliability of the results.

In addition to examining the positive effects of Internet use on the health of older adults and the underlying mechanisms involved, this paper seeks to address the limitations of previous research. To account for external influencing factors and resolve the issue of endogeneity between variables, this paper employs a difference-in-differences (DID) model. Furthermore, to address the challenge of small sample sizes, this study integrates six years of statistical data into pooled cross-sectional data. This approach not only increases the sample size but also allows for a more nuanced examination of the similarities and differences in the impact of Internet use on the physical health of older adults over different periods. Specifically, this study modifies the classic DID model to construct a pooled cross-sectional DID model, supplemented by a series of robustness tests to ensure the scientific validity of the conclusions.

Theoretical hypotheses

Social support theory emphasizes how interactions within social networks promote physical health by providing informational and instrumental support (Cohen & Wills, 1985). Older adults can access knowledge about disease prevention, healthy eating, and exercise through the Internet, and they can use health-monitoring devices or online medical services to enhance their ability to manage their health. At the same time, the core of self-determination theory lies in the influence of autonomy, competence, and relatedness on an individual's behavioural choices (Deci & Ryan, 2000). The Internet enables older adults to autonomously choose their health management methods, such as personalized exercise plans and dietary programs, thereby enhancing their sense of autonomy. The use of smart devices to monitor health data further strengthens their sense of

competence. By joining health-related interest communities, they can receive encouragement, which promotes the persistence of long-term healthy behaviour, ultimately contributing to improved physical health.

Hypothesis 1: internet use behaviour positively impacts the physical health of the older adults.

Hypothesis 2: internet use behaviour promotes the physical health of the older adults through the path of increasing exercise frequency.

Research design

Selection of External Impact Nodes

From a policy perspective, the ***Healthy China Plan 2030***, released by the Chinese government in 2016, serves as a top-level design for the health sector in China. It clearly states that the health issues of older adults should receive significant attention in the process of promoting national health. The policy particularly emphasizes the role of information technology as a vital tool in addressing the health challenges faced by older adults. From the perspective of Internet accessibility, official statistics indicate that in the second half of 2015, China's Internet penetration rate exceeded 50% for the first time (China Internet Network Information Center, 2016), marking the widespread adoption of digital technology across various sectors of society. Meanwhile, the ways in which older adults access information underwent significant changes during this period (as shown in Figure 1). Compared to traditional channels such as newspapers, magazines, and radio, the proportion of older adults accessing information via the Internet steadily increased, particularly between 2015 and 2017, when the proportion of older adults using the Internet as their primary information source saw a significant rise. Against this background, this paper identifies 2016 as a critical point when information technology began to exert an external influence on the information acquisition behaviour of older adults. It further explores the potential impact on older adults' health management and the implementation of related policies.

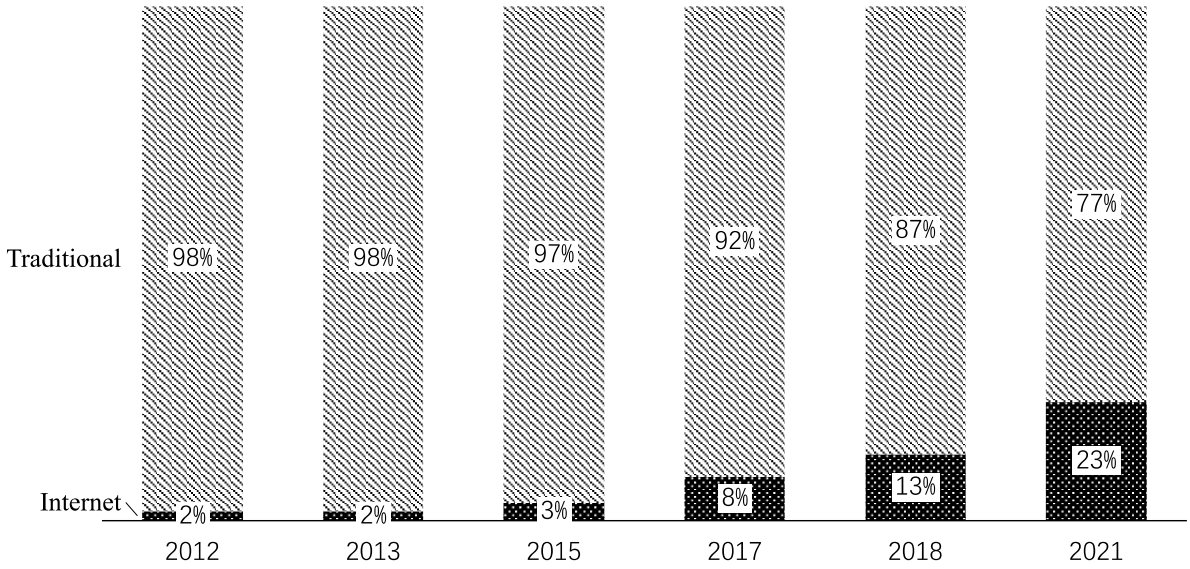


Figure 1. Information access channels for Chinese older adults (Source: The data are collected from CGSS database)

Construction of model

CGSS started in 2003 is a nationwide, comprehensive and continuous academic survey project by Renmin University of China. As one of the most important social surveys in China, CGSS has highly representative and widely applicable data. The CGSS data used in this paper are pooled cross-sectional data, observing different research subjects at different points in time. Therefore, this paper modify the classical DID model to construct a pooled cross-sectional DID model (Kiel & Mcclain, 1995; Yang et al., 2024). Formula (1) considers the time fixed effect based on the pooled cross-sectional DID model. Here, Yrepresents the dependent variable, the physical health of the older adults, with subscripts i and t representing the individual and the year, respectively. $treat_i \times post_t$ is the core explanatory variable, the interaction term between the post-policy dummy variable and the treatment group dummy variable. $treat_i$is the treatment group dummy variable, assigned a value of 1 for individuals with Internet use behavior, and 0 otherwise. Meanwhile, $post_t$is the time dummy variable. The external impact node selected in this paper is the year the ***Healthy China Plan*** (HCP) was issued, with post 2016 assigned a value of 1, and 0 otherwise. X_{it} represents other control variables, including individual and family-level variables. γ_t represents the time fixed effect. Since the data selected in this paper are pooled cross-sectional data, observing different individuals at different points in time, individual fixed effects are not considered. ε_{it} is the random disturbance term of the model. The estimated value of the coefficient. β_1 is the focus of this paper. If hypothesis 1 holds, the coefficient should be significantly positive.

$$Y_{it} = \beta_0 + \beta_1 treat_i \times post_t + \alpha X_{it} + \gamma_t + \varepsilon_{it} \qquad (1)$$

This study draws on data from the China General Social Survey (CGSS) for six selected years: 2012, 2013, 2015, 2017, 2018, and 2021, to construct the research model. Individuals aged 60 and above from each survey year are chosen as the study subjects, with samples containing outliers or missing values in key variables excluded from the analysis. The dependent variable is determined by the question: '*In the past week, has your health affected your ability to work*?' This variable is rated on a five-point scale: always, often, sometimes, rarely, and never, with corresponding values from 1 to 5. '*Always*' is assigned a value of 1, while '*Never*' is assigned a value of 5. The core explanatory variable, DID, represents the interaction between the post-policy dummy variable and the treatment group dummy variable. Control variables encompass individual and family-level characteristics, as outlined in Table 1.

Variables	N	Mean	Median	S.D.	Min	Max
Physical	21244	3.480	4	1.212	1	5
DID	21244	0.181	0	0.385	0	1
Gender	21244	1.503	2	0.500	1	2
Education	21244	3.537	3	2.432	1	13
Registration	21244	1.852	1	1.181	1	8
Matrimony	21244	3.999	3	1.759	1	7
Children	21244	2.549	2	1.484	0	24
Status	21244	4.176	4	1.769	1	10
Familyeco	21244	2.551	3	0.770	1	5

Table 1. Descriptive statistical table.

Empirical tests

Baseline regression

Column (1) of Table 2 demonstrates that Internet use exerts a significant positive effect on the physical health of older adults when control variables are not included. Once control variables are

introduced, the coefficient of the core explanatory variable decreases slightly but remains positive, highlighting the robustness of the positive influence of Internet use. In column (3), after accounting for time-fixed effects, the coefficient of the core explanatory variable continues to be significantly positive, with statistical significance at the 1% level. Overall, these results provide support for Hypothesis 1.

Variables	(1) Physical	(2) Physical	(3) Physical
DID	0.536***	0.260***	0.289***
	(26.24)	(12.16)	(11.70)
Gender		-0.122***	-0.121***
		(-7.49)	(-7.41)
Education		0.042***	0.042***
		(11.23)	(11.09)
Registration		0.068***	0.068***
		(9.20)	(9.22)
Matrimony		-0.010**	-0.010**
		(-2.04)	(-2.02)
Children		-0.080***	-0.080***
		(-13.64)	(-13.67)
Status		0.065***	0.065***
		(12.20)	(12.29)
Familyeco		0.228***	0.225***
		(18.37)	(18.02)
Constant	3.383***	2.728***	2.739***
	(369.39)	(59.85)	(55.45)
N	21,244	21,244	21,244
R-squared	0.029	0.119	0.121
Year FE	NO	NO	YES

Notes: (1) Robust t-statistics in parentheses; (2) *** p<0.01, ** p<0.05, * p<0.1

Table 2. Baseline regression results.

Robustness tests

Parallel trend test

This paper follows the methodology of Liu and Qiu (2016) to examine the trend of changes between the treatment and control groups. The results shown in Figure 2 indicate that, within the 99% confidence interval, all regression outcomes prior to the impact point are insignificant, suggesting that the treatment and control groups experienced similar trends before 2016. After the impact point, the physical health of older adults in the treatment group showed a significant improvement. However, five years later, in 2021, the reduced activity range of older adults due to the pandemic negatively impacted their health, causing the parallel trends test to fail for that year. Despite this, the sample used for model construction overall passes the parallel trend test

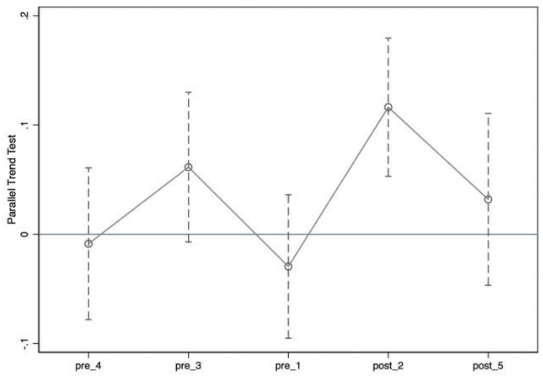


Figure 2. Parallel trend test results.

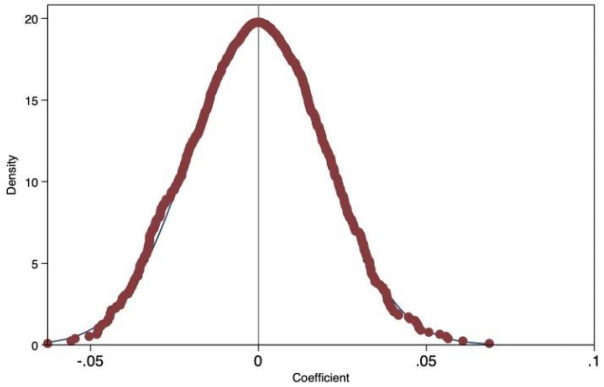


Figure 3. Placebo test results.

Placebo Test

To mitigate the potential influence of omitted variables on the estimation results, this paper adopts the approach of Cai (2016) , randomly selecting variables from the entire sample to form a pseudo treatment group, while assigning the remaining samples to a pseudo control group. Figure 3 presents the estimated coefficient kernel density, and the corresponding p-value distribution based on 1,000 random selections. The findings indicate that the average regression coefficient is close to zero, with the distribution of the treatment effect centered near zero, forming a symmetrical bell-shaped curve. These results suggest that omitted variables do not significantly bias the estimation, further confirming the robustness of the baseline regression outcomes.

PSM-DID

To address the issue of endogeneity in the sample, this paper employs propensity score matching (PSM) before conducting the difference-in-differences estimation. Following the approaches of Böckerman (2009) and Yi (2024), periodic propensity score matching is applied to the relevant data, using the same covariates for matching as the control variables in the baseline regression. The regression results are presented in Table 3. The first column of the table represents 1:1 nearest neighbor matching without replacement, while columns (2) through (4) show 1:1 to 1:3 nearest neighbor matching with replacement. Columns (5) to (7) present 1:1 to 1:3 caliper matching, and column (8) reports kernel matching. After performing propensity score matching to mitigate endogeneity, the core explanatory variables remain significantly positive, confirming the robustness of Hypothesis 1.

	(1) Nearest	(2) Nearest	(3) Nearest	(4) Nearest	(5) Caliper	(6) Caliper	(7) Caliper	(8) Kernel	(9) Repeated
DID	0.338***	0.325***	0.312***	0.302***	0.299***	0.307***	0.296***	0.288***	0.289***
	(6.76)	(7.86)	(9.48)	(10.11)	(9.49)	(10.42)	(10.52)	(11.63)	(12.00)
Constant	3.278***	2.985***	3.006***	3.033***	2.927***	2.986***	3.006***	2.753***	2.739***
	(31.57)	(27.73)	(35.23)	(39.26)	(34.09)	(38.30)	(40.98)	(55.14)	(58.00)
Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6,554	5,146	8,100	9,992	8,833	10,260	11,443	20,993	21,244
R-squared	0.117	0.080	0.087	0.089	0.087	0.090	0.091	0.119	0.121

Notes: (1) Robust t-statistics in parentheses; (2) *** p<0.01, ** p<0.05, * p<0.1

Table 3. PSM-DID and repeated random sampling.

Repeated random sampling

To address the issue of sample selection bias, this paper applies the Bootstrap method for repeated random sampling, following the approach of Li (2021), with a sample size of 10,000 and 500 random

extractions. The empirical results in column (4) of Table 3 show that the outcomes from repeated random sampling are largely consistent with the baseline regression results, further reinforcing the robustness of the conclusions.

Mechanism tests

Increase the frequency of exercise

The Internet can enhance the frequency of exercise among older adults, which in turn impacts their physical health. To examine this pathway, a mediation effect model is employed. Columns (1)-(3) of Table 4 present the results: (1) shows the baseline regression, (2) indicates that the variable representing the frequency of physical exercise in older adults is significantly positive, and (3) demonstrates that both the core explanatory variable and the mediation variable are significantly positive, with a reduction in the coefficient of the core explanatory variable, providing preliminary evidence for the existence of this pathway. Additionally, the Bootstrap test and Sobel test are conducted to validate the mediation effect. According to Table 5, the confidence interval for the indirect effect does not include zero, confirming the presence of the mediation effect. Moreover, the Sobel test returns a p-value of 0.001, significant at the 1% level, further establishing the mediation effect and supporting Hypothesis 2.

	(1) Physical	(2) Exercise	(3) Physical
DID	0.289***	0.410***	0.244***
	(11.70)	(11.66)	(9.97)
Exercise			0.112***
			(21.67)
Constant	2.739***	0.927***	2.632***
	(55.45)	(14.51)	(53.82)
Control	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	21,244	21,201	21,201
R-squared	0.121	0.181	0.141
Sobel Test (P Value)		0.000***	

Notes: (1) Robust t-statistics in parentheses; (2) *** p<0.01, ** p<0.05, * p<0.1

Table 4. Mediation effect results.

Effect	Coef.	Std. Err.	z	P>|z|	Normal Based [95% Conf. Interval]	
Indirect	0.055	0.004	12.990	0.000***	0.047	0.063
Direct	0.205	0.021	9.890	0.000***	0.165	0.246
Total	0.260	0.021	12.550	0.000***	0.220	0.301

Notes: (1) *** p<0.01, ** p<0.05, * p<0.1

Table 5. Bootstrap test result.

Heterogeneity analysis

Economic development heterogeneity

Based on the GDP levels of various regions in 2016, this paper divides the sample into economically developed areas and economically less developed areas. The results, presented in columns (1) and (2) of Table 6, indicate that the positive effect of Internet use on the physical health of older adults is more pronounced in economically developed regions compared to less developed ones. This difference may be attributed to the limited availability of exercise facilities and equipment for older adults in less developed areas.

	(1) Developed	(2) Less Developed	(3) Good	(4) Poor
DID	0.310***	0.227***	0.091***	0.318***
	(10.77)	(4.70)	(3.42)	(9.52)
Constant	2.874***	2.491***	4.037***	2.375***
	(48.76)	(27.74)	(70.24)	(39.93)
Control	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	14,627	6,581	8,215	13,019
R-squared	0.120	0.114	0.050	0.115
Inter-group Difference Test	0.071*		0.000***	

Notes: (1) Robust t-statistics in parentheses; (2) *** p<0.01, ** p<0.05, * p<0.1

Table 6. Heterogeneity analysis results.

Self- perceived health heterogeneity

Older adults exhibit varying perceptions of their own health. This section therefore examines how internet use affects physical health across different self-perceived health groups. Accordingly, the sample is divided into two groups: those with good self-perceived health and those with poor self-perceived health. The results, as shown in columns (3) and (4) of Table 6, indicate that internet use has a significantly stronger positive effect on the group with poor self-perceived health. These findings provide important insights for future policymaking: for older adults with chronic illnesses or disabilities, internet use proves to be an effective means of improving physical well-being.

Conclusion

With the rapid development of China's economy and the growing challenges posed by an aging population, physical health issues among older adults have become increasingly prominent. In this context, addressing these health concerns requires effective intervention measures. The expansion of Internet access offers a promising solution to this problem. This paper aims to examine the impact of Internet use on the physical health of older adults, providing a scientific foundation for the formulation of relevant policies and interventions. Grounded in social support theory and self-determination theory, the study employs a pooled cross-sectional difference-in-differences (DID) model, using data from the CGSS database spanning 2012 to 2021.

The findings indicate that Internet use has a significantly positive impact on the physical health of older adults, and the results remain robust. Furthermore, the mechanism analysis shows that Internet use enhances their physical health by increasing the frequency of exercise. The heterogeneity analysis reveals that Internet use has a stronger positive effect on older adults in economically developed regions and those with poor self-perceived health.

Future research should further investigate how different types of Internet usage behaviour affect the physical health of older adults to provide more comprehensive empirical evidence for policy development. Additionally, attention should be given to improving access to exercise facilities and equipment in economically less developed areas. Through continued research and policy refinement, this paper aims to further enhance the physical health of older adults and contribute to the realization of the Healthy China initiative. Such efforts will not only improve the quality of life for older adults but also significantly contribute to social harmony and stability.

Acknowledgements

This work is supported by the National Social Science Foundation of China (No.22&ZD327).

About the authors

Haoyuan Sun is a master student at the School of Information Management, Nanjing University, Nanjing, China. His research interests are in health informatics and digital humanities. He can be contacted at haoyuan.sun@smail.nju.edu.cn.

Zhenkang Fu is a Doctoral candidate at the School of Information Management, Nanjing University, Nanjing, China. His research interests are in technology and innovation management. He can be contacted at fuzhenkang@163.com.

Zhengtong Pu is a master student at the School of Information Management, Nanjing University, Nanjing, China. His research interests are in health informatics and digital humanities. He can be contacted at ztpu@smail.nju.edu.cn.

Qinghua Zhu is a Professor and Doctoral supervisor at the School of Information Management, Nanjing University, Nanjing, China. His research interests are in user information behavior, smart elderly care, and health informatics. He can be contacted at qhzhu@nju.edu.cn.

References

Ang, S., Lim, E., & Malhotra, R. (2021). Health-Related Difficulty in Internet Use Among Older Adults: Correlates and Mediation of Its Association With Quality of Life Through Social Support Networks. GERONTOLOGIST, 61(5), 693–702. https://doi.org/10.1093/geront/gnaa096

Böckerman, P., & Ilmakunnas, P. (2009). Unemployment and self-assessed health: Evidence from panel data. Health Economics, 18(2), 161–179. Scopus. https://doi.org/10.1002/hec.1361

Cai, X., Lu, Y., Wu, M., & Yu, L. (2016). Does environmental regulation drive away inbound foreign direct investment? Evidence from a quasi-natural experiment in China. Journal of Development Economics, 123, 73–85. https://doi.org/10.1016/j.jdeveco.2016.08.003

Cohen, S., & Wills, T. (1985). Stress, Social Support, and the Buffering Hypothesis. PSYCHOLOGICAL BULLETIN, 98(2), 310–357. https://doi.org/10.1037//0033-2909.98.2.310

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. PSYCHOLOGICAL INQUIRY, 11(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Gilleard, C., Hyde, M., & Higgs, P. (2007). Community and communication in the third age: The impact of Internet and cell phone use on attachment to place in later life in England. JOURNALS OF GERONTOLOGY SERIES B-PSYCHOLOGICAL SCIENCES AND SOCIAL SCIENCES, 62(4), S276–S283. https://doi.org/10.1093/geronb/62.4.S276

Guo, B., Zhang, X., Zhang, R., & Chen, G. (2022). The Association between Internet Use and Physical Exercise among Middle-Aged and Older Adults-Evidence from China. INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH, 19(24), 16401. https://doi.org/10.3390/ijerph192416401

Han, J., & Zhao, X. (2021). Impact of Internet Use on Multi-dimensional Health: An Empirical Study Based on CGSS 2017 Data. Frontiers in Public Health, 9. https://doi.org/10.3389/fpubh.2021.749816

Kiel, K., & Mcclain, K. (1995). House Prices During Siting Decision Stages—The Case of an Incinerator from Rumor Through Operation. JOURNAL OF ENVIRONMENTAL ECONOMICS AND MANAGEMENT, 28(2), 241–255. https://doi.org/10.1006/jeem.1995.1016

Li, Q. & Zhang, Y. (2021). Financial openness and resource allocation efficiency: Evidence from foreign banks entering China. Chinese Industrial Economy, 5, 95–113. https://doi.org/10.19581/j.cnki.ciejournal.2021.05.016

Liu, Q., & Qiu, L. D. (2016). Intermediate input imports and innovations: Evidence from Chinese firms' patent filings. Journal of International Economics, 103, 166–183. https://doi.org/10.1016/j.jinteco.2016.09.009

Sasaki, S., Sato, A., Tanabe, Y., Matsuoka, S., Adachi, A., Kayano, T., Yamazaki, H., Matsuno, Y., Nakano, A., & Watanabe, T. (2022). Internet use and physical activity of older adults during the COVID-19 pandemic: A cross-sectional study in a northern Japanese City. BMC GERIATRICS, 22(1), 688. https://doi.org/10.1186/s12877-022-03360-5

Yang, H., Zhang, Y., & Liu, L. (2024). Patent priority examination under mission orientation: Can speed and quality coexist? Technology in Society, 76, 102452. https://doi.org/10.1016/j.techsoc.2023.102452

Yi, Y., Zhang, L., Du, L., & Sun, H. (2024). Cross-regional integration of renewable energy and corporate carbon emissions: Evidence from China's cross-regional surplus renewable energy spot trading pilot. Energy Economics, 135, 107649. https://doi.org/10.1016/j.eneco.2024.107649

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Navigating end-of-life decision making: a thematic analysis of Uchinanchu American older adult's information behaviour

Kristina Shiroma
DOI: https://doi.org/10.47989/ir30iConf46900

Abstract

Introduction. End-of-life (EOL) involves complex decision-making, influenced by how individuals seek, share, and process information. This study examines the EOL decision-making information behavior of older Uchinanchu (Okinawan) Americans, an underrepresented ethnic minority.

Method. Qualitative, in-depth, semi-structured interviews were conducted with 18 Uchinanchu older adult participants to explore their information behavior in EOL decision-making.

Analysis. Reflexive thematic analysis, three key themes: (1) Prioritizing Family Well-being at EOL, where participants sought and shared information focused on minimizing burdens for adult children; (2) Keeping Cultural EOL Decision-making Customs, reflecting information exchange centred on filial responsibility; and (3) Navigating Fate when Making and Communication EOL Decisions, shaped by the concept of '*Shoganai*' (acceptance of fate).

Results. Participants information behavior revealed a nuanced balance between actively seeking information for decision-making and deferring to cultural norms or family guidance. The interplay between cultural values and EOL decisions influenced how information was gathered, shared, and applied within families.

Conclusion. These findings show that EOL information behavior is deeply shaped by cultural values and family dynamics, particularly in balancing individual preferences with collective responsibilities. Future research should examine how underrepresented populations engage in health-related information practices related to promote inclusivity in end-of-life healthcare practices and policies.

Introduction

Decision-making at end-of-life (EOL) involves complex, nuanced information processing and often includes multiple stakeholders. Older adults may need to gather and consider various information when considering sensitive decisions such as surrogate decision-makers, desired medical interventions, and the appropriate timing for these choices. However, many adults lack both formal advance care planning (ACP) documents and informal plans (Biondo et al., 2017; Dennis & Washington, 2018; Sharma et al., 2012). Older adults (aged 65 and older), in particular, are often hesitant to communicate EOL decisions with family or medical professionals (Lum et al., 2015). Similarly, caretakers and healthcare providers may feel uncomfortable discussion EOL plans with patients, leaving room for deferred decision-making (Scholz et al., 2020). This poses challenges in the U.S. healthcare system, where patient autonomy and proactive decision making are increasingly emphasized (Gao et al., 2015; Lau, 2023).

While medical EOL decision making often emphasises clinical guidelines and individual autonomy, cultural EOL decision-making incorporates broader social and familial values, which can either complement or contrast with Western medical recommendations. Cultural influences play a significant role in how individuals' approach and communicate their EOL choices. Different cultures have varying beliefs and practices surrounding death and dying, which can shape preferences for medical interventions, surrogate decision-makers, and communication about EOL wishes. For example, some cultures emphasize collective decision-making and family involvement, while others may prioritize individual autonomy or religious considerations. These cultural factors can influence whether individuals are more inclined to discuss their EOL preferences openly or defer such discussions to family members. Understanding these cultural nuances is crucial for tailoring EOL care to meet the needs of diverse populations.

Cultural perspectives of EOL decision making

Careful examination of EOL decision-making from a cultural perspective is increasingly important as the U.S. population ages and becomes more diverse (Gott et al., 2017). The Asian American population, one of the fastest-growing racial and ethnic groups in the country (Budiman, 2020), presents a range of unique characteristics, including diverse ethnic backgrounds, histories, and cultural approaches to health information seeking (Islam et al., 2016; Jang et al., 2018). Despite this diversity, there is limited understanding of the EOL decision-making preferences and behaviors of Asian American older adults, and even less is known about specific ethnic groups within this broader category.

Outside of EOL decision making, a recent few information behavior studies have extended information behavior theories to include marginalized communities, such as Chinese ethnic minority groups (Zhu & Liao, 2020) and speakers of indigenous languages (Burke, 2020). Zhu and Liao (2020) applied information theories to explore the information behavior of small ethnic minority populations in China. They found that information behavior was deeply rooted in cultural contexts, with participants primarily seeking information within their small world and only consulting external sources for urgent matters. However, there is little research focused on EOL decision making of Asian ethnic minorities, let alone from an information science approach (Shiroma et al., 2019).

Uchinanchu American older adults and their culture

Uchinanchu Americans are an ethnic minority group that are often aggregated under broader racial categories such as '*Asian*', '*Asian/Pacific islander*', '*Japanese*', and '*other*'. Uchinanchu and Uchinanchu Americans are descendants of the indigenous people of Uchina, more commonly known as Okinawa. The U.S Census American Community Survey (2022) estimates that between 2016 and 2020 the Okinawan American population was 3,526 people.

In Uchinanchu culture, EOL decision-making is deeply intertwined with both cultural traditions and medical considerations. Central to this is the Obon ritual, a three-day event that honors and celebrates ancestors, emphasizing the importance of family, community, and intergenerational connections. During Obon, the living engage in rituals such as preparing family altars, making food offerings, and lighting pathways to guide ancestral spirits back home, blending spiritual reverence with cultural practices. This ongoing relationship with the deceased, symbolized by the Eisa dance, reflects how Uchinanchu people view death not as an end but as a continuation of familial bonds (Johnson & Jaffe, 2008; Nelson, 2008). In this context, EOL decisions are not solely based on medical factors but are also shaped by cultural values—balancing respect for life with a deep sense of ancestral duty. Thus, medical decisions, while important, often coexist with cultural frameworks that guide how Uchinanchu individuals and families approach death and dying.

Beyond longevity studies (Buettner & Skemp, 2016), very little research focuses on Uchinanchu peoples' information preferences within the context of healthcare, decision making, and information science. This research gap, which promotes appropriation and commodification of a minoritized and/or indigenous culture's knowledge and behavior is problematic (Smith, 2012) and venerates an agenda of research being done to and on not with or for indigenous peoples (Smith, 2012).

To address this gap, this study aims to better understand the EOL decision-making information preferences and behaviours of Uchinanchu older adults by exploring three key research questions: (RQ1) What information do Uchinanchu older adults want to have when making EOL decisions? (RQ2) From what sources do Uchinanchu older adults draw on for the medical and/or cultural information they want when making EOL decisions? and (RQ3) What factors influence Uchinanchu older adults' information-seeking behavior in EOL decision-making? By focusing on Uchinanchu American perspectives, this study seeks to enhance understanding of diverse approaches to information behavior in EOL decision-making, contributing to more culturally sensitive healthcare practices and policies.

Method

Qualitative research methods are ideal for inductively examining information behaviours in EOL decision-making by allowing for participant's own words and experiences to drive the research and thus providing opportunities for ethically responsive and responsible research (K. L. Braun et al., 2014; V. Braun & Clarke, 2023; Smith, 2012). Therefore, to explore participants' in-depth perspectives, communication, and information behaviours in EOL decision-making, semi-structured, in-depth qualitative interviews were conducted.

Interview guide

The interview guide was developed with input from a literature review (Shiroma et al., 2019), results from a pilot study, and in consultation with community partners and health information behavior experts.

Participants

Eligibility criteria included: (1) self-identification as Uchinanchu (Okinawan) or Uchinanchu-American (Okinawan-American); (2) 65 years of age or older; and (3) proficiency in English or Hawaiian Creole English. A total of 18 older adults (aged 65 or older) participated in the study. Of these, 11 (61%) identified as female and 7 (39%) as male, with an average age of 74 years. Recruitment was facilitated by cultural community partners, supplemented by snowball sampling.

Interviews

Interviews were conducted in summer of 2023, with options for in-person or phone interviews based on participants' preferences. Interviews lasted 28-112 minutes, averaging 60-90 minutes. No participants discontinued the interview or refused to answer any question. Participants received

a $20 gift card for their participation. Audio recordings and transcriptions were managed using Otter.ai and QuickTime, with transcriptions manually reviewed for accuracy and then uploaded into Atlas.ti for coding.

Data analysis

Data analysis employed reflexive thematic analysis (RTA) (V. Braun & Clarke, 2022, 2023). RTA is an iterative process were the researcher expertise and personal experience guide theme identification. The analysis involved familiarizing with the data, coding based on a framework developed from EOL decision-making literature and cultural insights and generating themes through multiple iterations. The final themes were distinctly named and organized to reflect their contribution to the broader analysis, culminating in the report presented in this manuscript.

Ethical considerations

This study was approved by the Institutional Review Board at the University of Texas at Austin (#2019-09-0001). Approval and consultation from community partners ensured trust and accuracy in interview materials and final data. Informed consent was obtained from all participants prior to the interviews and recordings.

Analysis

Uchinanchu older adults' information preferences and behaviours in EOL decision making reflect a nuanced interplay between individual and familial considerations, shaped by cultural values and practices. Three themes were identified from the data: (1) prioritizing family well-being at EOL: burden minimization; (2) keeping cultural EOL decision-making customs: filial responsibility; and (3) navigating fate in EOL decision making: the concept of Shoganai.

Theme 1: Prioritizing Family Well-Being at EOL: Burden Minimization

The first theme, '*prioritising family well-being at EOL: burden minimization*' describes how Uchinanchu American older adults' approach EOL decision making with a strong emphasis on mitigating potential challenges that their EOL decisions might impose on their families, specifically their adult children. This focus on reducing burden encompassed financial, emotional, and physical aspects of EOL caregiving, reflecting a strategic approach to EOL planning that prioritizes family welfare. This motivation influenced their information seeking behaviours and decision-making processes, as they aimed to provide clarity and ease for their family members. For example, Participant 15 articulated the desire to pre-emptively address EOL decision to simply the process for her children:

> *I want my children to have..... I don't want them to have to make difficult decisions and guessing this and that what I wanted or what? They will just follow my wishes. And to make it easy for them* (P15).

This proactive stance is indicative of a deliberate effort to minimize the decision-making burden on family members by providing clear instructions and EOL decision-making information in advance. Participant 15 shared that by having EOL decisions made in advance, she believed her adult children would be able to easily make EOL decisions on her behalf.

Participant 9's approach to information seeking further illustrates this theme. Her EOL information behaviour included organizing a comprehensive EOL decision-making portfolio, including documentation on property, finances, and medical preferences, to aid her adult children: '*What else do I have in that binder? So, property, finance, medical…I don't want to be a burden.*' This information behaviour reflects the complex and multifaceted nature of EOL decision-making and underscore the participants' efforts to centralize and simplify access to critical information.

Similarly, Participant 6 discussed the importance of making financial EOL decision to avoid familial conflicts: '*It's [financial plans] so that they won't fight.*' The need to manage financial information in EOL decision making to prevent disputes underscores the broader objective of reducing emotional stress and friction within the family.

Participant 8's decision to include a Do Not Resuscitate (DNR) order highlights how the loss of independence can drive individuals to make EOL decisions that prioritize their autonomy and reduce the caregiving burden on others: '*I already made it. I said no resuscitation. If I can't take care of myself, aloha. And that's it*'. This communication of EOL preferences illustrates a strategic approach to information behaviour, where individuals proactively manage their EOL choices, formally or informally, to minimize the emotional and logistical burden on their families and caregivers.

Prior experience with EOL decision making for others has also influenced participant's information behaviours. Participant 11, reflecting on her past experiences with her husband's EOL, expressed the importance of making EOL decision to ease the burden on her family:

> *I don't want whoever's left to have to struggle. I want to make it as easy as possible for them. So, because I did for my husband before my husband died, I put together all the information I would need. On the occasion of his death, had all the phone numbers of people I would have to commute communicate with the organization I would have to communicate with that made it easy, because you don't want to have to think about this when somebody dies* (P11).

This quote underscores how past experiences with EOL decision-making for family members influence current information behaviour. Such prior experiences can motivate proactive information strategies in deliberate effort to streamline access to essential information, thereby mitigating emotional and logistical challenges faced by loved ones during a difficult time.

The theme of '*prioritizing family well-being at EOL: burden minimization*' reveals a critical aspect of Uchinanchu American older adults' EOL decision-making information behaviour. Participants' proactive strategies in organizing and communicating their EOL preferences reflect a concerted effort to minimize the impact on their family members, emphasizing the importance of clarity and preparation in the EOL decision-making process.

Theme 2: keeping cultural EOL decision-making customs: filial responsibility

The second theme of '*keeping cultural EOL decision-making customs: filial responsibility*' illustrates how cultural expectations and familiar roles shape information behaviour in EOL decision-making among Uchinanchu American older adults. This theme emphasizes that while participants might view EOL decision-making as a burden on their adult children, they themselves approach it with a strong sense of filial duty, particularly when it involves their parents.

While the eldest male adult child is often examined within the context of filial responsibility, Uchinanchu participants more frequently highlighted Uchinanchu women as central to EOL caretaking and decision making. These experiences of fulfilling filial responsibilities for their parents helped to shape participants' understanding of EOL decision-making practices and policies. Such responsibilities influenced participants' information behaviour by increasing their familiarity with medical EOL decision-making processes. For instance, Participant 11's role as a caretaker and EOL decision-maker for her parents exemplifies this dynamic:

> *They already had like a health directive. So, we knew we knew what to do. We didn't really have to make decisions because they had already like they had already had their funeral plan. They had their urns picked out. They had already what was to go on the urn, the inscription*

> *on the urn, and the Rinzai Zen emblem. So, we really didn't have to make decisions. Yeah, because they took care of everything* (P11).

Her experience with her parents' medical directives and funeral plans not only informed her understanding but also streamlined her own EOL decision-making process. This demonstrates how past EOL responsibilities shape the way individuals manage and navigate information related to EOL decisions.

Participant 5's positive experience with EOL decision-making was enhanced by the formalized information available through advance EOL decision-making forms, reflecting their commitment to filial responsibility. They noted, '*I helped with my mother-in-law's (EOL decision making). But it was simple, I just follow the instructions on the forms.*' This underscores how effective information management practices in EOL decision-making support the fulfillment of filial responsibilities, streamlining the process and reducing the burden on family members.

For Participant 2, the approach to his own EOL decision-making and caregiving was deeply influenced by the principle of filial responsibility, reflecting how he has internalized family obligations across generations. He described how he has adopted a family plan that mirrors his parents' approach to EOL care, explaining:

> *My dad, he's the oldest in family told me that it'd be my responsibility to take care of him when he gets older. And so, I accepted that. And when my brothers had his daughters, in case he passed away, we're in agreement that I would raise his kids. And when, if I were to pass away, my kids would be raised by my two sisters, with the financial backing of my brother* (P2).

This structured plan exemplifies the way filial duty shapes EOL information behavior by ensuring that family members are prepared to manage and support each other's EOL needs.

Currently, Participant 2 is engaged in caregiving for both his elderly spouse's mother and his grandchildren. He sees this responsibility as a manifestation of filial duty and anticipates reciprocal support from his children in the future, stating: '*Two generations, one up and two down. But, you know, that's what family do, take care of each other. We do our best and hopefully [adult children] will take some lessons.*' This ongoing caregiving role highlights how filial responsibility influences EOL decision-making and information behavior, demonstrating how cultural expectations inform the management of both current and future EOL responsibilities within the family.

Participants' experiences with their parents' EOL decision-making influenced their own information behaviours, reflecting a deep cultural commitment and sense of family obligation. This information behaviour illustrates the role of information in how participants manage EOL decision-making to align with cultural norms, including fulfilling their filial responsibilities and adhering to established familial roles.

Theme 3: navigating fate in EOL decision making: the concept of Shoganai

The third theme, '*navigating fate in EOL decision making: the concept of Shoganai*,' examines how participants approach EOL decision-making through the lens of Shoganai, a Japanese concept reflecting acceptance of one's fate. This theme reveals how the belief in accepting the inevitability of death influences participants' information behaviours regarding EOL decisions.

Participant 7 introduced the term Shoganai, explaining its meaning as '*the Japanese have a saying: Shoganai, [it] cannot be helped.*' He elaborated on how this mindset helps individuals cope with unpredictable events and death:

> *and some things, they're accidents. Some people's life is so much shorter, and this is unpredictable, I'll use that word, Shoganai. That's life, things happen, and we have no control over them.*

The concept of Shoganai influenced some participant's information behaviour by leading to a reluctance to engage in advance EOL planning. Participant 8 expressed a passive attitude toward death: '*I really don't think about it. Time comes, time comes.*' This mindset, which aligns with a cultural aversion to discussing death, often resulted in a lack of proactive EOL decision-making. Participant 8 further explained how having fewer responsibilities allowed them to avoid worrying about death: '*To me, life is just life. I'm not married, I don't have many responsibilities. If I die, I die. It's like live for today and don't worry about tomorrow.*' This mindset reveals how Shoganai can lead to minimal engagement with information related to EOL decisions, as it prioritizes acceptance of fate over planning.

In combination with a Shoganai mindset, the notion of having fewer responsibilities can diminish the urgency to consider EOL decisions, particularly if there is a pre-existing aversion to discussing or planning for death. Participant 6, who was adamant about not preparing for death, articulated the spontaneity of death and the perceived futility of planning for it:

> *Death is so spontaneous unless you're very sick. And you have a little bit more time with an individual. But when death comes quickly, you know, like you don't anticipate an accident in the car. And you die. There's no time to plan anything out* (P6).

Thus, the concept of Shoganai can overshadow the perceived value of EOL decision-making, as it promotes a 'live for the day' attitude.

However, for some participants, the inevitability and unpredictability of death prompted them to consider advance EOL plans. Participant 7 noted:

> *Sometimes life don't happen that way. You might have a stroke. You might have dementia, or you don't know what's going on. So, somebody still have to. You have to designate somebody to make the decisions for you* (P7).

Here, the Shoganai mindset helped them recognize the importance of making advance EOL decisions to manage uncontrollable health circumstances.

Maintaining a Shoganai mindset was also tied to reflections on what participants had learned from those who had already passed away. Participant 13, when reflecting on their experiences, invoked Shoganai by saying: '*Have your plans set out. I think mostly is to be grateful. And when the time comes, you know [it is] time to go move on.*' This view framed EOL information behaviour as a final act of gratitude, allowing for a transition with minimal worry, and emphasized the proactive approach in preparing for the future.

Conversely, Participant 12 discussed how dwelling on average lifespan statistics could lead to mental health challenges, particularly when not maintaining a Shoganai mindset:

> *You know, there are many people with psychological issues, mental issues. I mean, those that can be manipulated by saying, like I just said average lifespan is 78, I'm 60. And if they dwell on something like that, and you know, ticking bomb, oh I'm 77, I'm gonna die next year? I think some people think like that* (P12).

This highlights how mental health considerations can influence the Shoganai mindset, with some participants using it to alleviate worries and justify either avoiding or embracing EOL planning as a means of managing life's uncertainties.

Results

The results of this study provide a nuanced understanding of Uchinanchu older adults' information behaviour in navigating EOL decision-making through a small-world lens, revealing the intricate interplay between familial responsibilities, cultural values, and personal attitudes. The themes of mitigation of burden, filial responsibility, and the Shoganai mindset emerge as central to understanding this dynamic. These themes reflect how information behaviour in EOL decision-making is influenced not only by immediate family needs but also by broader cultural and generational contexts.

Mitigation of burden in EOL decision making

In Uchinanchu older adults' EOL decision-making, a key information behaviour observed is the strategic effort to mitigate the burden on their adult children. Participants actively sought and utilized information to develop plans that address potential emotional, financial, and logistical challenges their families might face. This forward-looking approach highlights how their information preferences are guided by a focus on the generational impact of their decisions. This finding aligns with previous research on the burdens encountered by family members during EOL planning (Gao et al., 2015; Lau, 2023). However, this study reveals a unique emphasis on practical concerns, such as family conflicts over property and wealth, that older adults proactively address.

Mitigation of burden for adult children is a significant theme in this context, reflecting a broader strategy to ease the potential impacts of their death on their descendants. Uchinanchu older adults are motivated to minimize practical, emotional, and financial burdens by providing clear and organized instructions for their family members. By anticipating and addressing potential challenges in advance, they aim to reduce stress and prevent familial discord. This proactive approach ensures their wishes are communicated and executed without placing unnecessary strain on their loved ones, highlighting the essential role of interconnected dynamics in fostering thoughtful and effective EOL planning.

This factor underscores how Uchinanchu older adults' information behaviour is shaped by the close-knit nature of their familial and cultural settings. The focus on burden mitigation reveals how familial ties and cultural expectations influence individuals to engage in meticulous planning, demonstrating a commitment to family cohesion and support. By considering the generational impact of their decisions, these adults navigate EOL planning with a deep awareness of their role within their family structure, thereby highlighting the interplay between practical concerns and cultural values in shaping their approach to EOL decision-making.

Filial responsibility in EOL decision making

Within these intimate, familial settings, the cultural expectation to care for one's parents drives individuals to approach EOL decisions with a strong sense of duty. This filial responsibility is often passed down through generations, shaping how older adults plan for their own EOL while considering the needs and expectations of their descendants. Filial responsibility in this context means that decisions are made not only to reflect personal preferences but also to honour familial obligations and streamline the caregiving process for family members, ensuring that all necessary arrangements are in place to uphold cultural and familial norms.

While filial duty is frequently emphasized in research on Asian American families (Gao et al., 2015; Lau, 2023), the findings of this study highlight that Uchinanchu older adults view this responsibility within a broader, multigenerational framework. Their EOL decision-making is influenced not only by the needs of their parents but also by a commitment to future generations, including children and grandchildren. This extended view of filial responsibility reflects a deep cultural value that prioritizes family cohesion and interdependence. Participants expressed a sense of duty that spans multiple generations, demonstrating that their information behaviours are intricately tied to maintaining family harmony and meeting cultural expectations across the familial lineage.

This multi-generational perspective is evident in how participants prepare for their own EOL, ensuring their decisions address both their immediate family's needs and their obligations to descendants. The concept of filial responsibility thus shapes their approach to EOL planning, integrating both the practical aspects of caregiving and the cultural imperatives of family support. This comprehensive view underscores the significance of family roles in shaping EOL decision-making processes, reflecting a broader cultural ethos that values both personal preparation and collective familial care.

Shoganai mindset in EOL decision making

The concept of Shoganai, or acceptance of fate, introduces a nuanced perspective on EOL decision-making within this population. Shoganai, which embraces the inevitability of death, influences how Uchinanchu older adults engage with planning processes. For some, this mindset leads to a more passive approach, reflecting a belief that certain aspects of life are beyond control and should not be overly planned or worried about. This contrasts with the fear-based avoidance often associated with death taboos in other Asian cultures (Biondo et al., 2017; Sharma et al., 2012), where discussions about death are frequently shrouded in anxiety.

Conversely, Shoganai also fosters a proactive approach to EOL planning, highlighting a duality where acceptance of fate coexists with the practical need to prepare for potential life events. Some participants embraced Shoganai by adopting a carefree attitude towards EOL planning, focusing on living in the present rather than worrying about the future. Others, however, used Shoganai to acknowledge life's unpredictability, prompting them to make provisions for unforeseen circumstances, such as designating decision-makers in case of incapacitation. This balance between cultural acceptance and practical readiness illustrates how Shoganai can shape both passive and active approaches to EOL planning, depending on individual interpretations.

Overall, while Shoganai might reflect a broader cultural hesitation around detailed EOL planning, all participants engaged with the process in some capacity. This suggests that Uchinanchu older adults navigate EOL decision-making with a balanced view, integrating their cultural values with practical considerations to manage what can and cannot be controlled.

Conclusion

This study explores how the end-of-life (EOL) decision-making of Uchinanchu older adults is profoundly shaped by the dynamics of their unique perspectives and cultural approaches, focusing specifically on the factors of burden mitigation, filial responsibility, and the Shoganai mindset. The proactive efforts to mitigate burdens reveal a strategic approach where individuals actively seek information to ease potential challenges for their descendants. Filial responsibility underscores the generational and cultural imperatives that shape how EOL decisions are made, demonstrating how familial obligations and cultural values guide information-seeking behaviors in a context of close-knit familial relationships. The Shoganai mindset introduces a nuanced perspective, balancing acceptance of fate with practical readiness, thus illustrating how cultural attitudes towards control and destiny influence both passive and active approaches to EOL planning.

These findings have significant implications for understanding how cultural and familial contexts shape information behaviors related to EOL decision-making. The emphasis on burden mitigation highlights the need for support systems that address the practical aspects of EOL planning, such as resources for managing financial and logistical challenges. The theme of filial responsibility suggests the importance of integrating cultural values into EOL planning tools and services, recognizing the role of multi-generational obligations. The dual nature of the Shoganai mindset points to the need for flexible approaches that accommodate varying attitudes towards fate and control.

This qualitative research is one approach by one researcher within one community at one specific point in time. As such, there are opportunities to extend our understanding of information behavior in EOL decision making. Future qualitative research could explore how these themes manifest in different cultural contexts and examine how those dynamics influence information behaviors in other settings. Additionally, further investigation into how the Shoganai mindset interacts with contemporary EOL planning practices for could provide deeper insights into the balance between cultural acceptance and practical preparedness for Uchinanchu American and other Asian American older adults. Future quantitative research may add empirical support and work towards generalizability. A longitudinal approach may help to develop insights on how EOL decision making changes over time. By expanding our understanding of these dynamics, future studies can enhance the development of culturally sensitive and practical EOL planning resources that better support individuals and families in navigating these critical decisions.

Acknowledgements

The author would like to acknowledge Dr. Bo Xie for her guidance as well as the support of this work through her Delores V. Sands Chair in Nursing Research Fellowship, School of Nursing at the University of Texas at Austin. Additionally, the author would like to gratefully acknowledge the participants of this study without whom this work would not be possible.

About the author

Kristina Shiroma is an Assistant Professor in the School of Information Studies at Louisiana State University, Baton Rouge, Louisiana, USA. She received her Ph.D. from the iSchool at the University of Texas at Austin. Kristina's research interests include aging, health, and cultural aspects of information. She can be contacted at kshiroma@lsu.edu.

References

Biondo, P. D., Kalia, R., Khan, R., Asghar, N., Banerjee, C., Boulton, D., Marlett, N., Shklarov, S., & Simon, J. E. (2017). Understanding advance care planning within the South Asian community. Health Expectations: An International Journal of Public Participation in Health Care and Health Policy, 20(5), 911–919. https://doi.org/10.1111/hex.12531

Braun, K. L., Browne, C. V., Ka'opua, L. S., Kim, B. J., & Mokuau, N. (2014). Research on Indigenous elders: From positivistic to decolonizing methodologies. The Gerontologist, 54(1), 117–126. https://doi.org/10.1093/geront/gnt067

Braun, V., & Clarke, V. (2022). Thematic analysis: A practical guide. SAGE.

Braun, V., & Clarke, V. (2023). Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. International Journal of Transgender Health, 24(1), 1–6. https://doi.org/10.1080/26895269.2022.2129597

Budiman, A. (2020). Key findings about U.S. immigrants. Pew Research Center. https://www.pewresearch.org/short-reads/2020/08/20/key-findings-about-u-s-immigrants/

Buettner, D., & Skemp, S. (2016). Blue Zones: Lessons from the world's longest lived. American Journal of Lifestyle Medicine, 10(5), 318–321. https://doi.org/10.1177/1559827616637066

Burke, M. (2020). Health information behavior of speakers of endangered languages. Journal of Librarianship and Information Science, 52(4), 1042–1049. https://doi.org/10.1177/0961000619899453

Dennis, M. K., & Washington, K. T. (2018). "Just let me go": End-of-life planning among Ojibwe elders. The Gerontologist, 58(2), 300–307. https://doi.org/10.1093/geront/gnw151

Gao, X., Sun, F., Ko, E., Kwak, J., & Shen, H.-W. (2015). Knowledge of advance directive and perceptions of end-of-life care in Chinese-American elders: The role of acculturation. Palliative & Supportive Care, 13(6), 1677–1684. https://doi.org/10.1017/S147895151500067X

Gott, M., Frey, R., Wiles, J., Rolleston, A., Teh, R., Moeke-Maxwell, T., & Kerse, N. (2017). End of life care preferences among people of advanced age: LiLACS NZ. BMC Palliative Care, 16(1), 76. https://doi.org/10.1186/s12904-017-0258-0

Islam, N. S., Patel, S., Wyatt, L. C., Sim, S.-C., Mukherjee-Ratnam, R., Chun, K., Desai, B., Tandon, S. D., Trinh-Shevrin, C., Pollack, H., & Kwon, S. C. (2016). Sources of health Information among select Asian American immigrant groups in New York City. Health Communication, 31(2), 207–216. https://doi.org/10.1080/10410236.2014.944332

Jang, Y., Yoon, J., & Park, N. S. (2018). Source of health information and unmet healthcare needs in Asian Americans. Journal of Health Communication, 23(7), 652–660. https://doi.org/10.1080/10810730.2018.1500660

Johnson, H., & Jaffe, J. C. (2008). Performing Japan contemporary expressions of cultural identity (1st ed.). Global Oriental.

Lau, S. S. (2023). Filial piety and U.S. family law: How cultural values influence caregiving, end-of-life, and estate planning decisions in Asian American families. Asian Pacific American Law Journal, 26(1). https://doi.org/10.5070/P326160687

Lum, H. D., Sudore, R. L., & Bekelman, D. B. (2015). Advance care planning in the elderly. The Medical Clinics of North America, 99(2), 391–403. https://doi.org/10.1016/j.mcna.2014.11.010

Nelson, C. T. (2008). Dancing with the dead: Memory, performance, and everyday life in postwar Okinawa. Duke University Press.

Scholz, B., Goncharov, L., Emmerich, N., Lu, V. N., Chapman, M., Clark, S. J., Wilson, T., Slade, D., & Mitchell, I. (2020). Clinicians' accounts of communication with patients in end-of-life care contexts: A systematic review. Patient Education and Counseling, 103(10), 1913–1921. https://doi.org/10.1016/j.pec.2020.06.033

Sharma, R. K., Khosla, N., Tulsky, J. A., & Carrese, J. A. (2012). Traditional expectations versus US realities: First- and second-generation Asian Indian perspectives on end-of-life care. Journal of General Internal Medicine, 27(3), 311–317. https://doi.org/10.1007/s11606-011-1890-7

Shiroma, K., Davis, N., & Xie, B. (2019). End-of-life information preferences of Asian older adults: A systematic review of the literature. Innovation in Aging, 3(Supplement_1), S412–S412. https://doi.org/10.1093/geroni/igz038.1536

Smith, L. T. (2012). Decolonizing methodologies: Research and Indigenous peoples.

U.S. Census Bureau. (2022). American Community Survey 5-Year Estimates, 2017-2021. U.S. Department of Commerce, Economic and Statistics Administration. Retrieved from https://data.census.gov

Zhu, M., & Liao, X. (2020). Chatman's theory of life in the round applied to the information seeking of small populations of ethnic minorities in China. Information Research: An International Electronic Journal. https://doi.org/10.47989/irpaper868

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

"You can't imagine how hard it is": hardship streamers in live crowdfunding

Yueru Yan, Pnina Fichman
DOI: https://doi.org/10.47989/ir30iConf47599

Abstract

Introduction. The combination of Live streaming and E-commerce also offers opportunities for people whose family members are severely ill, disabled, or old and facilitates their ability to raise funds by selling commodities and receiving virtual gifts. Unlike other streamers, hardship live streamers are driven by unique motivations and may employ distinct self-presentation techniques.

Method. We analysed TikTok video clips and sales data from 14 male and female hardship streamers.

Analysis. We used both quantitative and qualitative analyses of the data, including topic modelling, sentiment analysis, and thematic content analysis.

Results. We found that 1) hardship streamers used multi-modal cues to facilitate interactivity and authenticity, as they emphasized their hardship and promoted the product; 2) higher literacy level of the streamers and the use of a positive tone positively correlated with better sales; 3) hardship storytelling positively correlated with the time the audience stayed in their live streaming sessions; and 4) females focused more on their hardships, used a more negative tone, and talked faster than males.

Conclusions. We provide a unique understanding of the effectiveness of self-presentation strategies used by marginalized live streamers in raising funds on TikTok.

Introduction

Social live streaming services (SLSSs) are online platforms where users can broadcast real time content. China has witnessed rapid growth in its e-commerce markets and live streaming communities with the establishment of platforms like Douyin (the Chinese version of TikTok, hereafter referred to as TikTok) and Kuaishou (Tang et al., 2022). Alongside receiving virtual gifts from audience members, live streamers on these platforms can promote products in their broadcast rooms and in return, get a cut of the e-commerce revenue. This is known as *E-commerce live streaming* (ELS). In 2023, the ELS industry reached a market size of 4.9 trillion RMB (around 700 billion dollars) in China (IResearch, 2024).

And ELS has empowered a lot of marginalized population, like rural women (Tang et al., 2022), rural farmers (Yu and Zhang, 2022) and minority enterprisers (Li and Kang, 2021) in China. This study looks into another group of people who seek help online by engaging in ELS. In China, countless people and their families continue to face challenges in covering medical expenses that surpass the insurance coverage provided by the government (Huang et al., 2021). Some of them then turn to live streaming to earn money through the virtual gift received and the product sold. Online crowdfunding is defined as a process whereby people utilize internet platforms to solicit financial support from the public (Kim et al., 2018). Thus, we refer to live streaming to raise funds for a cause as *Live Crowdfunding* and live streamers seeking financial contributions by streaming about their life challenges as *hardship streamers*.

Many hardship streamers are elderly or come from rural backgrounds. However, much of the existing research on SLSSs has focused on the information behaviour of young, popular micro-celebrities or influencers (Fietkiewicz et al., 2018; Li et al., 2022; Lu et al., 2019; Zimmer & Scheibe, 2019), particularly on non-E-commerce topics such as video gaming, E-sports, and live performances. Regarding ELS, few studies have explored the strategies employed by streamers from marginalized groups or those facing significant challenges. Notably, there are significant differences in the self-presentation strategies used by traditional E-commerce live streamers and hardship streamers. For example, strong fan bases and opinion leadership typically associated with influencers—factors that affect purchase intentions (Farivar & Wang, 2022; Jegham & Bouzaabia, 2022)—may be absent in hardship streamers. Additionally, high-quality visuals and polished presentations that positively facilitate consumers' reactions to advertisements (Pozharliev et al., 2022) may be lacking in hardship live streams, because these streamers are often broadcasting in their rural houses. Understanding their technology usage is crucial to gain insight into its influence on marginalized groups and to better support their needs and further empower them with better design.

Previous research of online crowdfunding has typically focused on specific sites like GoFundMe and Kickstarter. These sites are characterized by a spatial and temporal separation between fundraisers and donors and an absence of significant non-verbal communication cues. However, SLSSs allow streamers to communicate and answer questions synchronically with their contributors; this immediacy enables nuanced expression and clarification, similar to face-to-face communication (Trevino et al., 1987). Contributors, in turn, consider the synchronous visual and verbal cues that the streamers express to make informed decisions on the spot. It is believed that more sophisticated communication mediums are more appropriate for the context of crowdfunding (Koch et al., 2015), and richer media can have a positive influence on decision quality (Kahai & Cooper, 2003). However, little research has looked at this specific type of online crowdfunding. Our study will explore the potential of *live crowdfunding* by investigating the effectiveness of self-presentation strategies used in synchronous communication by hardship streamers.

This study uses a mixed methodology to examine the multi-modal practices and self-presentation strategies used by hardship streamers on TikTok by analysing data from 14 streamers whose family

members are seriously ill, disabled, or old. Walther (1996) has found that impression management and selective self-presentation vary by gender, a claim also verified in the context of live streaming self-presentation (Freeman & Wohn, 2020; Wu et al., 2022). It is possible that self-presentation strategies of hardship live streamers will vary by gender as well, which may impact their crowdfunding effectiveness. Thus, we investigate the self-presentation strategies of 14 hardship live streamers, their variations by gender, and their effectiveness in crowdfunding by addressing the following questions:

- How do the hardship streamers present themselves?
- Do hardship streamers' presentations vary by gender?
- What self-presentation strategies of hardship streamers impact their effectiveness?

Data analysis revealed that hardship live streamers followed the same techniques used by other E-commerce live streamers when utilizing multi-modal cues in presentation, but they would emphasize their hardships when promoting products. Compared with males, females talk faster, use a more negative tone, and focus more on stories of personal hardship. There is evidence of a moderate correlation between hardship storytelling and audience watch time and a correlation between the emotion score and sales data.

Our research addresses a critical gap in the study of SLSSs by providing valuable insights into the diverse strategies marginalized streamers employ to present themselves online. These self-presentation practices help us better understand the empowering effects of ELS on marginalized communities in China. Furthermore, this study intersects social media practices with crowdfunding. The effectiveness of these strategies can inform the design of crowdfunding platforms and improve live streaming practices for fundraisers. Additionally, the research offers a fresh perspective on the information behaviour of rural and elderly populations—an area that remains underexplored in SLSSs studies. Ultimately, this research sheds light on the broader impact of live streaming on marginalized groups in non-Western contexts.

Literature review

E-commerce live streaming (ELS) in China

The rise of this new form of crowdfunding in China can be linked to the rapid expansion of the country's live streaming industry in recent years. Live streaming in China began in 2005 with users leveraging online chatrooms for indoor, performative broadcasts (Li & Lu, 2017), and it wasn't until 2015 the industry significantly extended its reach (Lu & Siegfried, 2021). Unlike in North America and Europe, China's live streaming industry is characterized by a higher prevalence of professional, full-time streamers who offer a more diverse array of content. Additionally, platforms have integrated more advanced features and linked with other online services (Lu et al., 2018).

However, previous studies from the perspective of streamers have mainly been conducted on video gaming live streaming platforms like Twitch (Freeman & Wohn, 2020) and Douyu (Gong et al., 2023; Wu et al., 2022) or popular chatting live streaming services, such as YouNow and Ustream (Fietkiewicz et al., 2018). These studies analyse the motivations (Fietkiewicz et al., 2018; Lu et al., 2019; Zimmer & Scheibe, 2019) and self-presentation (Freeman & Wohn, 2020; Wu et al., 2022) of the producers. They find that streamers adopt various self-presentation strategies to attract the audience (Wu et al., 2022), and that they make use of content outside the live streaming room to complement their self-presentation (Lu et al., 2019).

Driven by technological advancements such as 5G, virtual reality (VR), and artificial intelligence (AI) (Liu, 2020), ELS in China experienced significant growth and the market size reached 4.9 trillion CNY (around 700 billion USD) in 2023 (IResearch, 2024), which is about 10 times increase from 2019. The integration of live streaming and e-commerce allows consumers to purchase products directly from streamers within the live broadcast, facilitating immediate feedback and offering

deeper insights into consumer preferences (Lu & Siegfried, 2021). Some streamers on TikTok act as 'salesman' for the TikTok Shop, where companies put their product and set commission rates; streamers choose the products they want to sell and then put them into their live streaming session's cart to gain a portion of the sales. The ELS has notably empowered a lot of people in China, including ethnic minority startups (Li and Kang., 2021), rural women (Tang et al., 2022) and Intangible Cultural Heritage workers (Lu et al., 2019). Lu et al. (2019) focused on streamers who promote Intangible Cultural Heritage, revealing their altruistic motivation, and Li and Kang (2021) studied the ethnic minority startups flourishing on Chinese live streaming platforms, highlighting the diversity of participants in live streaming and their varied self-presentation.

Prior research on ELS focused on factors that influence sales performance. It is found that streamers/influencers play significant roles in the sales process (Zhang et al., 2022); thus, the self-presentation of streamers interests many researchers. Findings show that particular physical and linguistic characteristics are crucial for improving sales (Liao et al., 2023). Coherence between verbal and facial emotions promotes live sales (Yu et al., 2022), and consumers' purchase behaviour is affected by the product, social information contained in the streamer's speech, and emotional information (Chen et al., 2024). In ELS, female streamers often use discussions on everyday life to connect with customers, while male streamers tend to focus on popular topics to attract consumer attention (Wang & Jiang, 2021). This study aims to investigate the specific strategies used by hardship streamers in ELS and variations in gender.

Online crowdfunding

A growing number of people are using online medical crowdfunding platforms to request financial help from the public (Kim et al., 2018). The existing body of research on online crowdfunding tends to focus on asynchronous platforms tailored to charitable (Higueras et al., 2022; Sher & Su, 2019) or entrepreneurial objectives (Anderson & Saxton, 2016; Korzynski et al., 2021; Kromidha & Robson, 2016). Koch et al. (2015) found that sophisticated communication mediums are more appropriate in the context of crowdfunding: including video, pictures, and a well-formulated project description can increase success. Zhao et al. (2022) discovered that maintaining emotional consistency between visual and verbal modalities can alleviate challenges faced by online charity crowdfunding campaigns, with verbal signals being more impactful.

Many researchers have textually analyzed the narratives and project descriptions of web-based crowdfunding websites to investigate successful medical crowdfunding behaviour online (Hou et al., 2022). However, relatively few studies have used these methods to explore communication cues presented within live crowdfunding on short-video platforms. Previous research has examined the self-presentation strategies of individuals facing hardship, describing how they emphasize positivity and express their concerns anonymously (De Choudhury and De, 2014; Newman et al., 2011), and how they navigate expectations in asynchronous medical crowdfunding (Kim et al., 2018). Unlike traditional web-based fundraising platforms, live crowdfunding operates within a richer medium where fundraisers can talk with the audience in real time. This setting also offers a wider range of interactions, such as gifting, commenting, and e-commerce, potentially asking for varied self-presentation strategies from the fundraisers.

Methods

Data collection

To address our research questions, we used purposive sampling to choose eligible hardship streamers from TikTok in China (known as Douyin). First, we searched for 'hardship live streaming' ('苦难直播') and observed several live streaming rooms. Then, we identified more than 20 hardship live streamers and examined their profiles and live streams. We chose 14 of them, 7 male and 7 females, based on the following criteria: streamers that 1) stream at least once a week, as indicated by their streaming activities log and 2) sell commodities during their live streams. Since they are

all TikTok users, we were able to examine their profiles and the videos they previously posted to gain important contextual information, including the relationship between the fundraiser and the beneficiary. Typically, they pinned a video on their main page that addressed their hardship, offering further insights into the context of their fundraising efforts. As determined by the information they shared on their profiles or during their live streams, they are all from rural areas, and eight of them are over 60 years old (Table 1).

Streamer (Gender)	Beneficiary relationship to streamer	Talking pace (# of sentences/ second)	# of followers	RPM (yuan)	RPP (yuan)	Stay (seconds)	Viewers (# of viewers/ minute)
P1 (F)	Daughter-in-law	0.391	6436k	143.3	0.11	94	1859
P2 (F)	Grandchild, Self	0.118	5.77k	7.67	0.60	88	13
P3 (F)	Sons	0.197	635k	29.1	0.31	76	105
P4 (F)	Granddaughter	0.306	5k	20.05	1.15	41	14
P9 (F)	Daughter	0.270	50k	12.4	1.22	46	20
P10 (F)	Daughter	0.258	1242k	8.33	0.02	96	721
P11 (F)	Daughter	0.360	300k	144.5	0.78	67	113
P5 (M)	Mother	0.176	24k	2.68	0.01	24	136
P6 (M)	Son	0.262	17k	134.13	1.21	36	104
P7 (M)	Wife	0.158	49k	4.23	0.16	28	18
P8 (M)	Granddaughter	0.092	610k	272.6	0.42	71	726
P12 (M)	Son	0.272	70k	7.81	0.65	31	11
P13 (M)	Son	0.195	8k	80.8	1.25	53	47
P14 (M)	Sister	0.092	168k	4.26	0.02	95	1290

Table 1. Demographics, beneficiary relationships, and effectiveness of the selected streamers

Most streams lasted for over one hour, mainly between 8:00 PM and midnight in the streamers' time zone. Due to their streaming consistent patterns, we determined that video recordings of hour-long sessions would provide rich, suitable data. We recorded 45-60 minutes of live streaming from each of the 14 streamers, adjusted by their actual streaming length. Using iflyrec, which was designed by iFlytek (https://www.iflyrec.com), a well-known Chinese language processing company, we generated transcripts from each of the video clips. The first author checked the accuracy of each transcript manually, and the transcripts were translated by iFlytek into English to facilitate the second author's analysis and to ensure the study's reliability.

We collected the sales data and other live streaming-related information, including viewership and average viewing time, for all 14 streamers from kaogujia.com, a website that analyses data from *Douyin* (https://open.douyin.com). We used five measures to examine the effectiveness of live crowdfunding: the average number of viewers per minute in each recorded session ('Views'), the number of followers, the revenue earned per minute (RPM), the revenue per person (RPP); and the average audience stay time (Stay) for each session. The average revenue per streamer in our sample is 62.3 Yuan (approximately 9 US dollars) per minute (calculated based on the streaming session we recorded), the details are shown in Table 1.

Data analysis

First, we observed the live streaming sessions of the 14 streamers, alongside the featured products and the streamers' interactions with their audiences. We then performed a content analysis on the recorded videos to examine their multi-modal presentation. Next, we performed sentiment analysis, topic modelling, and thematic analysis on the recorded videos.

To develop a codebook for the multimodal cues and gain contextual information, we observed streaming sessions and joined some of the streamers' fan groups, focusing on 1) the reason for live streaming and the relationship between the streamer and the beneficiary; 2) the interaction between the streamer and the audience, especially the streamer's reactions to comments; 3) the

type of products sold alongside product selection and promotion; 4) the challenges of live streaming; and 5) the visual and audible multimodal cues utilized by participants.

Since research suggested that streamers' emotions could influence the sales and users' purchase intentions (Yu et al., 2022; Liao et al., 2023; Chen et al., 2023), we conducted sentiment analysis on the transcripts. We utilized Vader (Valence Aware Dictionary and Sentiment Reasoner; Hutto & Gilbert, 2014), which is a sentiment analysis tool designed for social media text, of the translated Chinese corpus. Then, we extracted the compound scores for each sentence. The compound score can range from -1 to 1, and it serves as an indicator of sentiment strength, where larger positive values represent stronger positive sentiments, and larger negative values indicate stronger negative sentiments (Hutto & Gilbert, 2014). We computed the mean of the compound scores across all sentences in each streamer's transcript as indicators of their final emotional tone.

Then, we conducted topic modelling and employed Jieba, a useful Chinese segmentation package in Python (https://github.com/fxsjy/jieba), for word segmentation. We manually added the words used frequently in the live streams into the dictionary when doing segmentation, then we calculated the TF-IDF (Term Frequency-Inverse Document Frequency) for each word. We selected the top 25 words for each streamer based on the TF-IDF score and generated the list of topics based on these words.

Supported by the topic modelling results (Table 2), we developed a code book for the thematic analysis (Table 3) with 6 codes. And then we sampled 4,200 sentences (300 sentences per streamers) to do the analysis.

Findings and discussion

How do hardship streamers present themselves?

We first describe the topics the streamers discuss during their live streaming sessions through the results of the topic modelling (word level of analysis), thematic analysis (sentence level of analysis), and multimodal cues (recorded video/streamer level of analysis). We end the section by discussing the relationships we found between the topics and the multimodal cues.

Based on our topic modelling, we found that the streamers portray a resilient image, sharing their struggles along with hopefulness. Based on the 25 most significant words identified by TF-IDF scores we identified the top six topics (Table 2).

	Topics	Example words	TF-IDF score
1	Welcoming	Welcome (‘欢迎’), good evening (‘晚上好’)…	0.065
2	Appreciation	Thanks, support, company…	0.164
3	Promoting	Link, Cart, Click on…	0.120
4	Interaction	Follow me, Give me a like…	0.051
5	Prospect	Come on, Hang on…	0.030
6	Storytelling	Grandma, Brother, Child…	0.082

Table 2. Top topics

The first topic involves welcoming the audience as they enter the streaming room. The second topic focuses on expressions of gratitude for audience purchases and support; this is a crucial aspect for fundraising streamers who rely heavily on audience goodwill for financial sustenance. The third topic involves introducing commodities in the streamer's virtual cart, which is called in Chinese the 'Little Yellow Cart' (‘小黄车’). The streamers introduce the products in the cart and encourage the audience to check them out. Most of the items are everyday essentials offered at relatively low prices, making them suitable for audience members to purchase. The fourth topic encourages audience interaction with the streamers, wherein streamers prompt actions such as

'Follow me' or 'Give me a like'. They also direct the audience to their accounts' main profile, where they post short videos explaining their hardship. To further facilitate ongoing interest in and financial backing of the streamers, they often display their bank account information on profiles and fan group chats. These actions also algin with previous studies where streamers would make use of content outside the live streaming room to complement their self-presentation (Lu et al., 2019).

The fifth includes expressions that demonstrate perseverance and hopefulness, aimed at uplifting both themselves and their audience, featuring motivational phrases like 'I will overcome it' or 'It will be better'. The sixth topic revolves around the hardships faced by the streamer or their family, in which streamers often invoke the names of family members, such as 'Grandma', 'Child', and 'Brother', most of whom are the beneficiaries of their fundraising efforts.

We analysed 4,200 sentences (300 sentences from each streamer) by assigning codes to each sentence and calculated the percentage of the themes in the data. We found that streamers most often promoted products, followed by hardship storytelling, and expressions of appreciation for audience participation and purchasing (Table 3).

	Themes	Description	Examples	Percentage[1]
1	Welcoming	Welcoming the audience who went into their streaming room.	Welcome to the room! (欢迎来到直播间), Good evening, everyone (大家晚上好).	0.037
2	Appreciation	Showing gratefulness for the audience's behaviours, like purchasing and staying in the room.	Thanks for the support and company (谢谢大家的陪伴和支持).	0.146
3	Promoting	Introducing the products for sale.	This is the extra-size trash bag (这是加厚加大的垃圾袋).	0.429
4	Interaction	Reading or answering audience comments; Encouraging the audience to follow or like their streaming.	Give me a like (点点赞); Please follow me (可以给我点个关注).	0.086
5	Prospect	Expressions of hopefulness.	I will overcome it. (会加油的); It will be better (会越来越好的).	0.024
6	Storytelling	Talking about their personal stories, focusing on their hardships.	This medicine needs to be taken for life. (药终身的); All I want is the child to live well (我只求孩子过得好). You can't imagine how hard it is. (你不知道有多难)	0.22

[1]percentage out of 4,200 sentences

Table 3. Prevalence of themes.

By observing, we get six multi-modal codes, as illustrated in Table 4. And then we count the number of the frequency of each code. Based on the analysis of the multimodal cues, we found that most of the hardship streamers used Mandarin and played hopeful music in the background. Most of the streamers 1) did not dress up nor design their background, opting instead a simple background; 2) used cues of their hardship in their usernames; and 3) promoted multiple products.

Cues	Code	Description	# of streamers
Audio	Language use	Streamers use of Mandarin (or dialect).	10
	Background music	Streamers use background music with hopeful melodies and sounds.	9
Visual	Retail type	Streamers promote multiple products (or single product).	9
	Hardship username	Username contains hardship cues (or not).	8
	Streaming background	Streaming background is simple (or well designed).	11
	Dressing	Streamers dress up during their live streaming (or not).	6

Table 4. Multimodal cues

Using the Mann-Whitney test, we found significant relationships between the multi-modal cues and the six themes. While most of the streamers spoke Mandarin (10 streamers), four streamers' dialects were evident; these four are all streamers over 60 years old. Streamers who used Mandarin put more emphasis on their interaction with the audience compared with those with a dialect; we found that language use is correlated with the Interaction theme (U=38.0, p=0.007). It is possible that those who spoke a dialect, were illiterate and thus unable to read the comments, barring them from interacting with their audience as much as the literate Mandarin-speaking streamers.

Nine streamers used background music to create a specific crowdfunding atmosphere during their live streams to enhance their presentation and potentially induce the audience's empathy with the streamer. Previous studies have found that live streaming with background music increased consumers' purchase intentions and arousal (Zhang et al., 2023). In the context of live crowdfunding, most of the streamers used soft and hopeful background music, including lyrics like 'Hello, tomorrow, smile with tears' ('明天你好，含着泪微笑'), 'My love protects you like an angel' ('我的爱像天使守护你'). Interestingly, streamers who used background music interact with the audience more; the usage of background music was related to the Interaction theme (U=41.0, p=0.008).

Two retailing models are common on TikTok; the first occurs when streamers collaborate directly with specific companies, and the other is when streamers choose any products that are available on the TikTok Shop and add them into their yellow cart to promote said products. Evidence shows that selling more products and spending more time on each product in a live stream drive both gross merchandise volume (GMV) and fan growth (Chen et al., 2023). Nine of the 14 hardship streamers did not have collaboration with specific companies while five of them focused on one or two products. Promoting one product allowed streamers to have a more centralized focus., while promoting multiple products allowed the streamers to appeal to wider audiences.

Eight of the 14 hardship streamers used the name of and their relationship to the beneficiary in their streaming username; at times, they added words to their usernames that indicated their hardship or their current condition. Their usernames were, for example, 'XX's father is the bravest' or 'The disabled XX'. Usernames are an important visual code that helps the audience understand the purpose of live crowdfunding. This strategy is useful for allowing streamers to attract their audience by indicating their purpose up front.

A streamer's background influences consumers' purchase intentions, and the visual design that encourages consumers' positive emotions and purchase involves moderate visual complexity (Shang et al., 2022; Tong et al., 2022). Eleven of the 14 streamers streamed from their home using a simple or plain background, while only three of them designed their background to appear more professional. We found that the designed background is correlated with the theme of product promotion (U=0.0, p= 0.006). It is possible that streamers who designed their background live stream similarly to other e-commerce streamers; a designed background and heavy promotion of

products and can intentionally lead to their larger economic gain. Streamers with a random or empty background were less engaged in promoting products in their live streaming.

Lastly, while professional e-commerce streamers dress up and put obvious effort into steaming by projecting an appealing physical appearance (Gao et al., 2021), less than half of the streamers in our sample dressed up, most opting instead to dress casually. The audience estimates the streamers' trustworthiness based on simple cues such as the streamer's appearance, which, alongside their perception of e-commerce, impacts the audience's purchasing decisions (Sun et al., 2019). However, hardship live streamers' crowdfunding differs, because it is possible that a less appealing appearance can increase the audience's perception of their trustworthiness; just like the rural female streamers who also promote products online to make a living (Tang et al., 2022), they demonstrated their authenticity mainly through showcasing local traditions, living spaces and landscapes.

In summation, our findings demonstrate how hardship live streamers use multi-modal cues during their ELS by adopting an authentic, grateful, and resilient image through a combination of hardship storytelling and production promotion. They take advantage of the richness of live streaming that helps them share more visual and verbal cues to provide and receive immediate feedback from their audience (Daft & Lengel, 1986). Compared with other E-commerce live streamers, the hardship live streamers in our study try to differentiate themselves by highlighting their unique circumstances.

The streamers in our sample use visual, audible, and verbal (textual) cues to present an authentic and comprehensive self-image as grateful, resilient, and polite individuals in need of help. ELS streamers play a key role in establishing the audience's trust in the products through their interactivity and authenticity (Greenwald, 2020). Hardship streamers aim to establish their audience's trust in their authenticity and their hardships. To do this, streamers play soft and encouraging music, have simple backgrounds, and wear modest outfits. Some streamers speak with a dialect and stream from their simple, unfurnished house, which helps to establish their authenticity. They differentiate themselves from other popular E-commerce live streamers, who stream with decorated backgrounds (Tong et al., 2022) and utilize appealing visual presentations. Incorporating elements of their hardship into their usernames to provide immediate context for viewers is another aspect of their self-presentation that differ from other ELS streamers.

Do hardship streamers presentations vary by gender?

We did not find gender differences in the multi-modal cues, but we did find that females demonstrate a faster talking pace, use more negative tones, and emphasize their hardships more often than male streamers do.

We found that females tend to use a more negative tone compared to males (Table 5); the sentiment score generated from Vader for females is significantly lower than for male streamers ($U=9.0$, $p=0.028$). Female streamers tend to have a faster-talking pace, in terms of the number of sentences per second, than males ($U=39.0$, $p=0.037$). Female live streamers have significantly longer audience retention (Stay) length ($U=15.0$, $p=0.030$) than males, in terms of the average viewing time per audience. Importantly, RPP and RPM did not vary based on streamer gender (Table 5).

	Talking pace	Sentiment score	RPM	RPP	Stay (s)
Female	0.271	0.120	52.193	0.600	72.571
Male	0.178	0.208	72.358	0.531	48.286
Mann-Whitney U	39.0	9.0	18.0	20.5	15.0
P-value	0.037	0.028	0.222	0.327	0.030

Table 5. Variations by Gender

Analysis of the 6 themes by gender revealed no gender differences in any of themes, except for one: female hardship streamers are significantly more inclined to discuss their hardships (storytelling) in live streaming (Table 6). Like other female e-commerce live streamers who tend to talk more about their everyday life (Wang & Gong, 2017; Wang & Jiang, 2021; Wang & Liu, 2019), the female hardship streamers also talked more about their personal life to effectively appeal to the audience.

	Welcoming	Appreciation	Promoting	Interaction	Prospect	Storytelling
Female	0.019	0.119	0.384	0.065	0.017	**0.351**
Male	0.054	0.172	0.474	0.106	0.031	**0.089**
Mann-Whitney U	21.5	17.0	20.0	19.0	20.0	6.0
P-value	0.373	0.186	0.305	0.264	0.301	0.011

Table 6. Theme Frequency by Gender

In summary, female and male hardship live streamers adopt similar visual and verbal cues in their live streaming. They differ in the pace of their speech, their speaking tone, and the extent of their hardship storytelling. Hardship streamer's effectiveness, in terms of RPM and RPP, is not significantly different based on gender, but audiences tend to stay significantly longer on female live streaming sessions.

What self-presentation strategies of hardship streamers impact their effectiveness?

In this session, we begin by exploring the correlations among effectiveness measurements, followed by an examination of the relationship between emotion scores and these measurements. Finally, we talk about how multi-modal cues relate to effectiveness metrics.

The results of the Pearson correlation between the five measures of effectiveness with the sentiment scores and the six themes are shown in Table 7. As shown in Table 7, we found that the number of views per minute (Views) is positively correlated with the time the audience stayed (Stay) and the money they spent per minute (RPM). The revenue of the fundraiser (RPM) is positively correlated with the number of followers (Followers) and the average number of audience members per minute (Views). However, the longer the audience stayed in the streaming session (Stay) does not relate to a higher revenue per minute (RPM). It may be results of that the variation of the number of followers in our samples and interests in the live streaming content doesn't necessarily contribute to the purchase. They may contribute in other ways like gifting.

We also found a positive correlation between the emotion score and sales data (RPP), and a negative correlation between the emotion score and the time the audience stayed in the live streaming room. Expression of hardship (Storytelling) is correlated with the number of followers and with the time the audience spent in the live streaming room (Stay). Mentioning hardship through storytelling increased the time the audience stayed in the live streaming room and was correlated with a higher number of followers. While accentuating their hardship struggles helps

the hardship live streamers achieve their fundraising objectives, we found that it can negatively influence the streamer's emotion presented, which may adversely affect sales. However, the streamers' expressions of gratitude and hopefulness help mitigate this negative impact. Instead of keeping talking their hardship while streaming, nearly all streamers maintain a grateful and resilient image, frequently expressing appreciation to their audience for purchases, staying in the room, or showing goodwill in the comments. They also convey hopefulness for the future to keep the emotional tone positive, positively contributing to the consumers' purchase intention, which is consistent with previous study in ELS. Therefore, it is suggested that fundraisers balance product promotion and hardship storytelling to maintain a positive atmosphere, which would be beneficial for sales.

	Views	Followers	Stay	RPP	RPM	Sentiment score	Welcoming	Appreciation	Promoting	Interaction	Prospect	Storytelling
Views	1.00	0.79	**0.67***	**-0.54***	**0.54***	-0.50	-0.42	-0.51	-0.27	0.37	0.21	0.38
Followers		1.00	0.47	-0.34	**0.60***	-0.45	-0.28	-0.43	-0.18	-0.15	-0.20	**0.54***
Stay			1.00	-0.41	0.25	**-0.51***	-0.64	-0.28	-0.19	0.36	0.18	**0.30***
RPP				1.00	-0.03	**0.22***	0.14	0.16	0.14	-0.45	-0.06	0.13
RPM					1.00	0.07	-0.21	-0.43	0.22	-0.18	-0.15	0.16
						1.00	0.68	0.63	0.42	-0.21	0.01	-0.75

* $p < 0.05$

Table 7. Correlation matrix

To examine the relationship between the multimodal cues (visual and audio) and the effectiveness of streamers, we used the Mann-Whitney test. We found that 1) language use is related to RPP: streamers who used Mandarin made more money (RPP) than the streamers using dialect (U=29.5, p=0.050); and 2) streamers who focused on promoting multiple products kept the audience in their streaming session for longer (Stay) (U=41.0, p=0.008), but did not generate higher revenues (either RPP or RPM). It is possible that the use of Mandarin correlates with another explanatory variable that may or may not differ from other streamers, but we did not examine that in this study and hope that future research may uncover it.

Though we don't find significant correlation between the theme of interaction and sales performance examined in this study, it would potentially improve fundraising outcomes in a long term. Taking advantage of the interaction features like following, gifting, and liking, the hardship streamers encouraged the audience to interact with them by suggesting liking their videos, following their account, and join their fan groups. These strategies help turn the audience's short-term, one-time contribution into long-term, potentially continuous contributions.

Limitations

One of the limitations of the study is its small sample size, and given the sampling method, the hardship streamers in the sample may not represent hardship streamers as a whole. Hardship streamers from other countries or those who face other challenges may present different strategies and may exhibit other behavioural patterns; future research may examine larger and more heterogeneous samples. In addition, we analysed the live streaming session, but a better understanding of the motivations and intentions of the streamers calls for in-depth interviews of these streamers as they engage in live crowdfunding; future research may utilize other methods to address these types of research questions.

Conclusions

Through an analysis of 14 male and female hardship streamers engaged in ELS, this study examined how these individuals leveraged technology to seek online assistance and empowerment through fundraising live streams on the Chinese TikTok. Streamers employed visual, audible, and textual cues to project an authentic, grateful, and resilient persona, particularly when focusing on hardship narratives and product promotion. While they followed the same techniques used by other E-commerce live streamers when utilizing multi-modal cues in presentation, they distinguished themselves from other ELS streamers by highlighting their struggles through strategic use of background settings, music, and content. The study found a moderate relationship between the portrayal of hardship and the length of time audiences spent in live streaming sessions. However, rather than continually emphasizing hardship, maintaining a positive atmosphere in ELS was found to be more effective for fundraising success. While gender of the streamers didn't play a role in their fundraising success, there were gender differences in their streams: females had a faster talking pace, used a more negative tone, and focused more on personal hardship storytelling than their male counterparts. Overall, this study makes a significant contribution to the literature by illustrating how ELS can empower marginalized populations and identifying potential strategies to enhance this empowerment.

Acknowledgements

We would like to thank Cori Kemper for proofreading an earlier version of the manuscript.

About the authors

Yueru Yan is a PhD student at the Information and Library Science at Indiana University of Bloomington. Her research interests include computational social science, social media, and machine learning. She could be contacted via yueryan@iu.edu.

Pnina Fichman is a Professor of Information Science in the Luddy School of Informatics, Computing, and Engineering and the director of the Rob Kling Center for Social Informatics. Her research in social informatics mainly focuses on online trolling. She could be contacted via fichman@iu.edu.

References

Anderson, K., & Saxton, G. D. (2016). Smiles, babies, and status symbols: The persuasive effects of images in small-entrepreneur crowdfunding requests (SSRN Scholarly Paper 3670385). https://papers.ssrn.com/abstract=3670385

Burtch G, Chan J. Investigating the relationship between medical crowdfunding and personal bankruptcy in the United States: evidence of a digital divide. MIS Quart 2019 Jan 1;43(1):237-262.

Chen, J., Luo, J., & Zhou, T. (2024). Research on Determinants Affecting Users' Impulsive Purchase Intention in Live Streaming from the Perspective of Perceived Live Streamers' Ability. *Behavioural Sciences*, 14(3), 190.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. Management Science, 32(5), 554–571. https://doi.org/10.1287/mnsc.32.5.554

De Choudhury, M., & De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 71–80. https://doi.org/10.1609/icwsm.v8i1.14526

Farivar, S., & Wang, F. (2022). Effective influencer marketing: A social identity perspective. *Journal of Retailing and Consumer Services*, 67, 103026.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding (arXiv:2007.01852). arXiv. http://arxiv.org/abs/2007.01852

Fietkiewicz, K. J., Dorsch, I., Scheibe, K., Zimmer, F., & Stock, W. G. (2018). Dreaming of stardom and money: Micro-celebrities and influencers on live streaming services. Social Computing and Social Media. User Experience and Behaviour: 10th International Conference, SCSM 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I 10, 240–253.

Freeman, G., & Wohn, D. Y. (2020). Streaming your identity: Navigating the presentation of gender and sexuality through live streaming. Computer Supported Cooperative Work (CSCW), 29(6), 795–825. https://doi.org/10.1007/s10606-020-09386-w

Gao, X., Xu, X. Y., Tayyab, S. M. U., & Li, Q. (2021). How the live streaming commerce viewers process the persuasive message: An ELM perspective and the moderating effect of mindfulness. Electronic Commerce Research and Applications, 49, 101087.

Gong, M., Bao, X., & Wagner, C. (2023). Why viewers send paid gifts: The role of social influence on massively multiplayer online games live streaming. *Information Processing & Management*, 60(4), 103371.

Greenwald, M. (2020). Live streaming e-commerce is the rage in China. Is the US next?. Forbes, available at: https://www.forbes.com/sites/michellegreenwald/2020/12/10/live-streaming-e-commerce-is-the-rage-in-china-is-the-us-next/(accessed 14 October 2021).

Heath S. Does medical crowdfunding perpetuate affordability problems? Patient Engagement HIT. 2020 Oct 28.

Hou, X., Wu, T., Chen, Z., & Zhou, L. (2022). Success Factors of Medical Crowdfunding Campaigns: Systematic Review. Journal of Medical Internet Research, 24(3), e30189. https://doi.org/10.2196/30189

Huang, J., Shen, H., Cao, Q., Cai, L., & Cheng, X. (2021, May). How medical crowdfunding helps people? A large-scale case study on the Waterdrop fundraising. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15, pp. 220-229).

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

IRESEARCH. (2024). 2023 China Live E-commerce Industry Research Report [2023 年中国直播电商行业研究报告]

Jegham, S., & Bouzaabia, R. (2022). Fashion influencers on Instagram: Determinants and impact of opinion leadership on female millennial followers. *Journal of Consumer Behaviour*, *21*(5), 1002-1017.

Kahai, S. S. and R. B. Cooper (2003). "Exploring the Core Concepts of Media Richness Theory: The Impact of Cue Multiplicity and Feedback Immediacy on Decision Quality." Journal of Management Information Systems 20 (1), 263–299.

Kenworthy, N., Dong, Z., Montgomery, A., Fuller, E., & Berliner, L. (2020). A cross-sectional study of social inequities in medical crowdfunding campaigns in the United States. PLOS ONE, 15(3), e0229760. https://doi.org/10.1371/journal.pone.0229760

Kim, J. G., Hong, H., & Karahalios, K. (2018). Understanding identity presentation in medical crowdfunding. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3173574.3173708

Koch, J. A., & Siering, M. (2015). Crowdfunding success factors: The characteristics of successfully funded projects on crowdfunding platforms.

Li, L., & Kang, K. (2021). Why ethnic minority groups' online-startups are booming in China's tight cultural ecosystem? Journal of Entrepreneurship in Emerging Economies, 15(2), 278–300. https://doi.org/10.1108/JEEE-08-2021-0322

Lin, J., & Lu, Z. (2017). The rise and proliferation of live-streaming in China: insights and lessons. In HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part II 19 (pp. 632-637). Springer International Publishing.

Liao, M., Fang, J., Han, L., Wen, L., Zheng, Q., & Xia, G. (2023). Boosting eCommerce sales with livestreaming in B2B marketplace: A perspective on live streamers' competencies. Journal of Business Research, 167, 114167.

Lu, Z., Annett, M., Fan, M., & Wigdor, D. (2019). "I feel it is my responsibility to stream" Streaming and engaging with intangible cultural heritage through livestreaming. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–14.

Lu, Z., Xia, H., Heo, S., & Wigdor, D. (2018). You watch, you give, and you engage: a study of live streaming practices in China. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-13).

Magnuson, M. J., & Dundes, L. (2008). Gender differences in "social portraits" reflected in MySpace profiles. CyberPsychology & Behaviour, 11(2), 239-241.

Newman, M. W., Lauterbach, D., Munson, S. A., Resnick, P., & Morris, M. E. (2011). It's not that i don't have problems, i'm just not putting them on facebook: Challenges and opportunities in using online social networks for health. CSCW 2011.

Pozharliev, R., Rossi, D., & De Angelis, M. (2022). A picture says more than a thousand words: Using consumer neuroscience to study instagram users' responses to influencer advertising. *Psychology & Marketing*, 39(7), 1336-1349.

Saleh, S. N., Ajufo, E., Lehmann, C. U., & Medford, R. J. (2020). Crowdfunding medical care: a comparison of online medical fundraising in Canada, the United Kingdom, and the United States. MedRxiv, 2020-03.

Shang, Q., Ma, H., Wang, C., & Gao, L. (2023). Effects of background fitting of ELS on consumers' purchase intentions: a cognitive-affective perspective. Psychology Research and Behaviour Management, 149-168.

Sher, S. T.-H., & Su, N. M. (2019). Speedrunning for charity: How donations gather around a live streamed couch. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 48:1-48:26. https://doi.org/10.1145/3359150

Tang, N., Tao, L., Wen, B., & Lu, Z. (2022). Dare to Dream, Dare to Livestream: How E-Commerce Livestreaming Empowers Chinese Rural Women. CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3491102.3517634

Tong, X., Chen, Y., Zhou, S., & Yang, S. (2022). How background visual complexity influences purchase intention in live streaming: The mediating role of emotion and the moderating role of gender. Journal of Retailing and Consumer Services, 67, 103031.

Trevino, L. K., Lengel, R. H. and R. L. Daft (1987). "Media Symbolism, Media Richness, and Media Choice in Organizations: A Symbolic Interactionist Perspective." Communication Research 14 (5), 553–574.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. Communication Research, 23, 3–43.

Wang, M., & Jiang, T. (2021). Gender language in live streaming from the perspective of sociolinguistics. Years of Youth, 19, 23–24.

Wang, L., & Gong, H. (2017). Onlookers of female anchors: The game of power in live streaming. Journal of China Women's University, 5, 47–51. https://doi.org/10.13277/j.cnki.jcwu.2017.05.007

Wang, T., & Liu, Q. (2019). Self-construction and identity of young women in network live streaming. Comtemporary Youth Research, 4, 97–103.

Wang, Y., Lu, Z., Cao, P., Chu, J., Wang, H., & Wattenhofer, R. (2022). How live streaming changes shopping decisions in E-commerce: A study of live streaming commerce. Computer Supported Cooperative Work (CSCW), 31(4), 701-729.

Wang, Y., Xu, Y., & Dong, C. (2023). Manipulating "Beauty": Beautification Technology and Its Relational Applications on Chinese Livestreaming Platforms. Journal of Chinese Women's Studies, 4, 116–128.

Wu, Y., Li, Y., & Gui, X. (2022). "I am concerned, but…": Streamers' privacy concerns and strategies in live streaming information disclosure. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), 1–31.

Xu, J., Zhang, R., Guo, J., & Gong, J. (2022). Bullet screen and discipline: Gender gaze in webcast based on Douyu big data. Chinese Journal of Journalism & Communication, 146-151+153-167. https://doi.org/10.13495/j.cnki.cjjc.2022.04.003

Yu, M., Chen, H., & Dou, Y. (2023). Selling with Emotion: How Livestream Influencers' Facial and Verbal Expressions Drive Mobile Commerce Sales. Available at SSRN 4141860.

Yu, Z., & Zhang, K. (2022). The determinants of purchase intention on agricultural products via public-interest live streaming for farmers during COVID-19 pandemic. *Sustainability*, 14(21), 13921.

Zhao, K., Zhou, L., & Zhao, X. (2022). Multi-modal emotion expression and online charity crowdfunding success. Decision Support Systems, 163, 113842.

Zimmer, F., & Scheibe, K. (2019). What drives streamers? Users' characteristics and motivations on social live streaming services

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Motivational duality and fake news reporting behavior: a polynomial regression with response surface analysis

Xiao-Liang Shen, Lin-Yao Liu, and Yang-Jun Li
DOI: https://doi.org/10.47989/ir30iConf47596

Abstract

Context. Promoting social media users to report fake news is a crucial for curbing its spread, yet active reporters remain limited compared to those contributing to its proliferation.

Purpose. This study examines how users' moral motivations affect their fake news reporting behavior. Specifically, it investigates how the dual dimensions of moral motivation (i.e., agency and communion), both in congruence or incongruence, affect reporting behavior.

Method. Grounded in the reconciliation model of moral centrality, this study develops research hypotheses and tests them using over 13,000 reports and more than 4 million posts from 15,256 social media users. Empirical analysis employs dictionary-based text analysis method, and polynomial regression with response surface analysis.

Findings. Users with congruent agency and communion motivations are more likely to report fake news, and this likelihood increases as the strength of these motivations grows. User heterogeneity analysis shows that unverified users report more when communion exceeds agency, while verified users report more when agency exceeds communion.

Originality. This study presents a novel perspective on fake news reporting behavior by linking it to moral motivation and examining how the congruence or incongruence between agency and communion shapes reporting behavior, with distinct effects for verified and unverified social media users.

Introduction

The proliferation of fake news on social media platforms has caused numerous harms and has become an urgent issue requiring immediate attention (Kankanamge et al., 2025; Liu et al., 2023; Wang et al., 2021). Particularly with the development of artificial intelligence and deepfake technologies, public concerns about fake news have intensified (Gupta et al., 2024; Jiang et al., 2024). In response, internet giants such as Facebook, TikTok, and X have gradually implemented systems for users to report fake news. These reporting systems aim to harness the '*wisdom of the crowd*,' encouraging the public to take a more active role in detecting and reporting fake news (Shen and Wu, 2024; Wang et al., 2022; Wu et al., 2023). However, the reality is that the number of users who actively report fake news is limited, especially when compared to the vast number of users contributing to its spread. Existing research calls for the adoption of more effective intervention strategies to curb the spread of fake news (Kim et al., 2019; Li et al., 2022). In response to these calls, this study seeks to explore how to encourage fake news reporting behaviors from the perspective of moral motivation.

Moral motivation refers to the process by which individuals make moral judgments and are motivated to act in accordance with their moral perspective, often characterized as the willingness to do what is right (Sun et al., 2024). Given that reporting fake news is a prosocial behavior that reflects an individual's sense of social responsibility and moral duty, moral motivation may offer an explanation for users' more proactive engagement in such actions. It could serve as a key factor driving social media users to report fake news. However, to the best of our knowledge, our current understanding of whether and how moral motivation influences social media users to report fake news remains very limited.

Existing research in social psychology posits that moral motivation comprises two dimensions: agency and communion (Walker and Frimer, 2007). Agency refers to self-enhancing motives, characterized by the pursuit of power, status, achievement, or dominance. Communion, on the other hand, refers to other-enhancing motives, manifested in a universal concern for the disadvantaged, strangers, or the ecological environment (Frimer et al., 2012). It is commonly believed that self-interested agency motivation is immoral, as it hinders moral behavior, while other-oriented communion motivation is viewed as moral as it promotes moral behavior (Zhao et al., 2020). However, recent studies challenge this view, suggesting that agency is morally neutral, with its effect on moral behavior contingent upon whether the agency motives align with communal goals (Bell, 2024). Therefore, there remains controversy regarding how agency and communion influence individual behavior. This paper seeks to address the existing conflict in moral decision-making and explore how the agency and communion dimensions of moral motivation shape the behavior of reporting fake news. In line with this, our study proposes the following research question: How do agency and communion dimensions of moral motivation jointly influence social media users' fake news reporting behaviors?

Another limitation of existing research lies in the simplistic approach of measuring moral motivation through subtractive terms ($communion - agency$) or divisive terms ($communion / agency$), without treating agency and communion as independent variables (Zhang and Yu, 2018). These approaches exhibit the following flows: combining agency and communion into a single score implicitly assumes that both agency and communion contribute equally to the positive and negative aspects of behavior. Notably, such practices fail to reveal the distinct effects of agency and communion on user behavior, thereby undermining the prerequisite for elucidating the mechanisms of influence, that is, maintaining the independence of agency and communion. Furthermore, the existing analyses cannot determine whether the outcome variable is related to both factors or just one of them. From a methodological perspective, it is necessary to treat agency and communion as two independent variables for measurement and analysis. Polynomial regression with response surface analysis (PR-RSA) offers a promising solution by allowing the

independent examination of agency and communion within the data analysis process (Li et al., 2024; Liang et al., 2022). This method overcomes the operational limitations inherent in prior studies and provides an opportunity to reveal the mechanisms through which the dimensions of moral motives, agency and communion, influence the behavior of reporting fake news.

This study explores how the moral motivational duality affects fake news reporting behavior using PR-RSA. Drawing on the reconciliation model of moral centrality, we propose three hypotheses regarding the impact of agency and communion on fake news reporting behavior. We collected field data from Weibo, including over 13,000 reports from 15,256 reporters and more than 4 million posts from their homepages. As a major social media platform in China, Weibo is similar to X platform in terms of its widespread use and influence. Using PR-RSA method, we examine how the changes in fake news reporting behavior—whether increasing, decreasing, or remains unchanged—are influenced by the congruence or incongruence between agency and communion. Additionally, we explore the distinct roles that agency and communion play in promoting fake news reporting behavior for verified versus unverified social media users.

Literature review and theoretical background

Antecedents of online reporting behavior

Existing studies have examined online reporting behaviors across various contexts, including fake news (Wang et al., 2022), hate speech (Kim et al., 2023), and online harassment (Wong et al., 2021). These studies primarily examined social, platform, content, or individual-level antecedents of online reporting behavior. At the social level, existing literature focuses on how the user's reporting behavior is affected by the gender of users who post counterspeech and the popularity of these counterspeech in the social environment (Kim et al., 2023; Wu, 2024). At the platform level, researchers have studied how the design of the interfaces or reporting systems affects user reporting behaviors (Gimpel et al., 2021; Wilhelm et al., 2020; Zhou et al., 2023). For instance, Gimpel et al. (2021) observed that the combined application of injunctive and descriptive social norms within the user interface leads to the most substantial improvements in fake news reporting behavior. At the content level, reporting behavior varies when encountering different online content formats and characteristics (Wang et al., 2022; Wilhelm et al., 2020). For example, Wang et al. (2022) proposed that the presence of videos in fake news posts can increase the number of social media users reporting them to social media platforms. At the individual level, existing literature primarily focuses on how users' perceptions influence reporting behavior from the perspective of individual cognition (Naderer et al., 2023; Rodriguez Müller et al., 2024; Wong et al., 2021). For example, Wong et al. (2021) found that factors such as perceived responsibility, perceived emergency, perceived self-efficacy, perceived anonymity, and perceived outcome effectiveness significantly impact users' willingness to report social media harassment. However, prior studies have greatly overlooked the perspective of moral motivation, which is crucial for a deeper understanding of online reporting behavior.

Considering that online reporting behavior is a form of prosocial behavior reflecting individuals' judgments of '*right*' and '*wrong*', individuals with strong moral motivations are more likely to engage in such behaviors. From this perspective, our research differs from prior studies in that it seeks to understand how social media users' moral motivations influence fake news reporting behavior. Investigating the underlying moral motivation can help uncover the deeper reasons behind users' online reporting behaviors.

The reconciliation model of moral centrality

Early research on how the agency and communion dimensions of moral motives influence moral behavior has presented two competing viewpoints: the interference hypothesis (Bellah et al., 2007) and the synergy hypothesis (Lapsley, 2004). Both perspectives agree that communion is moral, but they differ their views on agency. The interference hypothesis posits that agency is anti-moral,

while the synergy hypothesis views agency as morally neutral. To address the above academic debate, Frimer and Walker (2009) integrated both the interference and synergy hypothesis concerning the psychological functions of moral exemplars and proposed a reconciliation model of moral centrality. This model hypothesizes that the relationship between moral concerns and self-interest (i.e., communion and agency) can transform from one of mutual competition to one of reconciling.

The key to understanding the reconciliation model lies in the concept of moral centrality, which refers to the extent to which morality is central to an individual's identity (Frimer and Walker, 2009). Identity does not pertain to an individual's intrinsic nature, but rather to an external orientation (Taylor, 1989). This orientation includes the things and activities that an individual considers '*good*, *worthwhile*, *admirable*, *or valuable*', and thus identifies with. In other words, identity represents the image an individual seeks to project to others and be recognized for, and it is inherently malleable (Flanagan, 2009).

Research indicates that identity often involves the pursuit of one's self-interests, which can out-compete other motives, such as concern for the plight of others (Bauer and Wayment, 2008). In this sense, when an individual's behavior is aimed at shaping their own image, with a focus on how others perceive their actions and how these actions contribute to the evaluation of their identity, the underlying motivation is centred on personal benefit. Therefore, if an individual's behavior is contingent on their identity, it can be considered motivated by agency. The idea provides a framework for integrating identity (agency) and morality (communion). When morality becomes central to an individual's identity (i.e., when engaging in moral behaviors contributes to shaping one's identity), there is an opportunity for the effective integration of communion and agency. Consequently, moral behavior can be seen as driven by enlightened self-interest (Frimer and Walker, 2009). The reconciliation model of moral centrality challenges the dichotomy between agency and communion, bridging the gap between moral judgment and moral behavior, known as the judgment-behavior gap. It also suggests that coordinating agency motivation and communion motivation will maximize the occurrence of moral behavior.

Hypotheses development

Previous studies suggest that even when individuals judge an action to be morally right, the motivational power of this judgment to spur action is relatively weak, illustrating a judgment-action gap (Walker and Frimer, 2007). Adhering to ethical norms alone is insufficient to motivate individuals to take action. From a developmental perspective, the reconciliation of agency and communion signifies moral maturity, gradually evolving towards a motivational system that closely aligns with moral exemplars (Frimer and Walker, 2009). In other words, only when social media users integrate self-interests with the public good—viewing the achievement of communal goals as beneficial to personal values—it greatly enhances their fake news reporting behaviors. Sanders et al. (2018) also point out that,

> *when being a moral person is not important to the sense of self, moral drivers are less likely to translate into heightened motivation to act selflessly and into actual ethical behavior* (p. 633).

Therefore, the driving force behind moral behavior is not merely about others. It can and should also involve self-concern (Walker, 2013). On social media platforms, reporting fake news can curtail its spread, thereby protecting other users from its influence. The communion motivation (caring for others) is evidently essential for encouraging the behavior of reporting fake news, but the agency motivation (concern for self-interest) is also important and should not be viewed as a hindrance. The primary reason is that when social media users experience a lack of congruence between agency and communion, perceiving reporting fake news as beneficial to others but not offering personal gains, such as enhancing their social reputation, they will be less likely to be

motivated to report fake news. Thus, we contend that, compared to social media users whose agency and communion are congruent, those whose motivations are incongruent are less likely to report fake news. Accordingly, we hypothesize:

Hypothesis 1. The incongruence between agency and communion in moral motivation is negatively related to social media users' fake news reporting behavior.

The integration of agency and communion motives among social media users signifies an advanced development of moral function. However, under the premise that users can fully reconcile agency and communion motives, how does the synchronized variation in the strength of these two motives affect the behavior of reporting fake news? Previous research indicates that, in addition to the structure of moral motivation (where moral exemplars can integrate agency and communion motives), the strength of moral motivation also explains moral behavior (Frimer et al., 2011, 2012; Walker, 2013). At the strength level, existing research has found that moral exemplars exhibit greater agency and communion motivations than comparison ordinary individuals (Frimer et al., 2011). In other words, when users are better able to reconcile agency and communion, especially when this reconciliation occurs at a higher level, it increases the likelihood of adopting the corresponding behavior. In the context of this study, under the condition that users' agency and communion motivations are well reconciled, stronger agency and communion motivations will promote users to report more fake news. Conversely, when users' agency and communion motivations are relatively weaker, their willingness to report fake news may be lower, resulting in fewer reports of fake news. Therefore, as the congruence between agency and communion motives increase from low to high, fake news reporting behavior will also increase. Thus, we hypothesize:

Hypothesis 2. Fake news reporting behavior will be higher when congruence between agency and communion occurs at high levels, compared to when it occurs at lower levels.

The reconciliation model suggests that if moral behaviors contain elements that can enhance self-worth—specifically, if moral behaviors can legitimately be seen as self-enhancing and self-promoting—individuals will have a stronger motivation to act morally (Walker, 2013). This is because helping others may benefit the helper, especially when such behaviors are noticed by observers, which can establish a positive reputation and lead to personal benefits (Frimer et al., 2014). Social media platforms have two distinct user groups: verified and unverified users, and they derive different personal benefits from engaging in moral behaviors. Unverified users (i.e., anonymous users) conceal their real identities, presenting a virtual persona instead. In contrast, verified users' real identities (such as their name and affiliation) must be authenticated by the platform and made publicly available. Therefore, compared to unverified users, verified users are more likely to enhance their personal reputation and gain additional benefits by engaging in moral actions, such as reporting fake news. Based on this, we argue that unverified users are more likely to report fake news when their communion motivation outweighs their agency motivation (i.e., when they emphasize other-enhancing motives). On the other hand, for verified users, due to the additional personal benefits they may obtain, verified users with a motivation system where their agency motivation is higher than their communion motivation (i.e., when they focus more on self-enhancing motives) are more likely to report fake news. Accordingly, we hypothesize:

Hypothesis 3a. For unverified social media users, when their moral motivation of communion is higher than agency, they will report more fake news.

Hypothesis 3b. For verified social media users, when their moral motivation of agency is higher than communion, they will report more fake news.

Research setting

Context

To test our research hypothesis empirically, we leveraged data from Weibo, a prominent and widely used social media platform in China. Users can not only post, like, share or comment on posts but also report fake news to the platform by clicking a button next to the post and selecting the reason '*fake news*' for reporting. Given the transparency of Weibo's reporting system, along with its resemblance to reporting systems found on other social media platforms like Facebook and X platform, Weibo presents an ideal environment for our study. In particular, the platform provides details regarding reported records, including the user who reported the post (reporter), the time of the reports, and evidence that supports or disproves the authenticity of the post. This feature enables us to retrieve the number of report times made by reporters based on unique homepage links, and provides access to their homepages to collect demographic information and posting history. More explanations can be found below.

Data collection and sample

We collected our data for the first time in January 2023 and further updated it in March 2024. We gathered a total of 13,346 report records. A single report record contains only one instance of fake news; however, that instance of fake news may be reported by multiple users. Each report of fake news by the reporter (referring to social media users who report fake news) will be considered a valid report. Since one piece of fake news can be reported by multiple users, meaning that one report record may involve multiple valid reports, we have collected a total of 24,853 valid reports involving 16,629 reporters as of March 2024. Next, we focused on these reporters and went to their homepages to collect their posting history. Due to instances of reporters deactivating their accounts and the participation of official institutional accounts in reporting, we identified 15,256 reporters and collected a total of over 4 million posts from their personal homepages. Ultimately, our dataset includes the number of times each reporter participated in reporting, along with their detailed demographic information such as gender, verification status, credit level, membership level and registration time, as well as their posting history. The posting history mainly refers to the content and timestamps of each post.

Variable operationalization

The dependent variable, $\ln(ReportedTimes + 1)_i$, represents the logarithmic transformation of the number of reported times made by reporter i. We have two independent variables: $Agency_i$ and $Communion_i$, which respectively represent the scores of the reporter i on the dimensions of agency and communion in moral motivation. In the acquisition of scores for agency and communion, word frequency analysis is a commonly used quantitative method because the utilization of words reflects the semantic dimensions of psychological constructs (Tausczik and Pennebaker, 2010). Specifically, we employed the moral motivation dictionary to measure the $Agency_i$ and $Communion_i$ of reporters, which was originally designed by Frimer for LIWC. It contained of 349 words for the agency dimension and 146 words for the communion dimension, and it has been extensively used by scholars in the field of library and information science. To facilitate its application across diverse backgrounds, the moral motivation dictionary used in related studies was translated into simplified Chinese, with Cronbach's alphas of .95 and .79 for agency and communion words, respectively (Zhang and Yu, 2018). After a series of translations and tests of dictionary validity, the latest Chinese version of moral motivation dictionary obtained from the original authors retains 722 agency words and 290 communion words. We calculated the agency and communion dimensions in moral motivation for each reporter using the following formula:

$$Agency_i = \frac{\sum Freq_{A,j}(T_i)}{722 * |T_i|},\ Communion_i = \frac{\sum Freq_{C,j}(T_i)}{290 * |T_i|} \quad (1)$$

where, $Agency_i$ and $Communion_i$ represent the agency and communion motivation scores for reporter i, respectively; T_i represents the set of all posts posted by reporter i; $Freq_{A,j}(T_i)$ and $Freq_{C,j}(T_i)$ denote the frequency of agency or communion word j in all posts posted by reporter i, respectively. Additionally, we also incorporated a set of control variables in our estimations, as shown in Table 1.

Variables	Description
Ln (ReportTimes+1)$_i$	Log of the number of reported times made by reporter i
Agency$_i$	The score obtained by reporter i on the agency dimension
Communion$_i$	The score obtained by reporter i on the communion dimension
Gender$_i$	1 if reporter i is male; 0 otherwise
Ln (DaysSinceRegist)$_i$	Log of the number of days between registration and data collection date
Verification$_i$	1 if reporter i verified; 0 otherwise
VIP$_i$	The actual membership level of reporter i, with level 1 recorded as 1, and so on.
Credit$_i$	Divided into four levels, with reporter i categorized as low level = 1, moderate level = 2, good level = 3, and excellent level = 4
Ln (NumberOfPosts)$_i$	Log of the number of all posts posted by reporter i

Table 1. Summary for the key variables.

Analysis and results

Analytical model

We used the PR-RSA to analyse and provide empirical evidence for hypothesis. As recommended by Edwards and Parry (1993), we zero-centred $Agency_i$ and $Communion_i$ prior to calculating the second-order terms to avoid multicollinearity and facilitate the interpretation of the results. Specifically, a polynomial regression analysis was conducted on users' fake news reporting behavior, as illustrated below:

$$\ln(ReportedTimes)_i = b_0 + b_1 Agency_i + b_2 Communion_i + b_3 {Agency_i}^2 + b_4 Agency_i \times Communion_i + b_5 {Communion_i}^2 + controls + \varepsilon_i \quad (2)$$

Where i = 1, 2, ..., 15256; $b_0, b_1, \ldots, b_5$ are model parameters to be estimated; and ε_i is the error term.

While polynomial regression is useful in estimating the relationship between dependent and independent variables, the estimated coefficients are challenging to interpret. Therefore, we adopted response surface analysis. The coefficients of polynomial regression are used to construct the response surface, which provides a three-dimensional visualization to reveal the relationship between the variables. In which, $Agency$ and $Communion$ are plotted on the perpendicular horizontal axes, while $\ln(ReportedTimes + 1)$ is plotted on the vertical axis. The features of the response surface relevant to this study include the slopes and curvatures of the congruence incongruence line. The following are the basic equations used in the PR-RSA.

$$a_1 = b_1 + b_2$$
$$a_2 = b_3 + b_4 + b_5;$$
$$t_1 = (b_1 + b_2)/\left(\sqrt{SE_{b1}^2 + SE_{b2}^2 + 2cov(b_1 b_2)}\right);$$
$$t_2 = (b_3 + b_4 + b_5)/\left(\sqrt{SE_{b3}^2 + SE_{b4}^2 + SE_{b5}^2 + 2cov(b_3 b_4) + 2cov(b_3 b_5) + 2cov(b_4 b_5)}\right) \quad (3)$$

$$a_3 = b_1 - b_2$$
$$a_4 = b_3 - b_4 + b_5;$$
$$t_3 = (b_1 - b_2)/\left(\sqrt{SE_{b1}^2 + SE_{b2}^2 - 2cov(b_1 b_2)}\right);$$
$$t_4 = (b_3 - b_4 + b_5)/\left(\sqrt{SE_{b3}^2 + SE_{b4}^2 + SE_{b5}^2 - 2cov(b_3 b_4) + 2cov(b_3 b_5) - 2cov(b_3 b_5)}\right) \quad (4)$$

$$P_0 = (b_2 b_4 - 2b_1 b_5)/(4b_3 b_5 - b_4^2)$$
$$E_0 = (b_1 b_4 - 2b_2 b_3)/(4b_3 b_5 - b_4^2)$$
$$p_{11} = \left[b_5 - b_3 + \sqrt{(b_3 - b_5)^2 + b_4^2}\right]/(b_4)$$
$$p_{10} = E_0 - p_{11} P_0 \quad (5)$$

Where a_1 and a_2 are the slope and curvature of the congruence line, respectively, and their significance levels are tested by t_1 and t_2. Similarly, a_3 and a_4 represent the slope and curvature of the incongruence line, respectively, with their significance levels tested by t_3 and t_4. P_{11} and P_{10} are the slope and intercept of the first principal axis, respectively, with their significance levels are tested through bootstrapping. The coefficients $b_1, b_2, \ldots, b_5$ are standardized coefficients estimated using equation (2), where SE is the standard error of the coefficient, and cov is the covariance between the coefficients.

As for how to interpret the meaning of these features, first, the slope (a_1) of the surface along congruence line represents how the reconciliation between agency and communion relates to reporting behavior. It shows variation of reporting behavior across the continuum from low to high when the strengths of agency and communion are the same. The curvature (a_2) of the surface along this line indicates whether the relationship between the strength of integration and reporting behavior is curvilinear. Second, the curvature (a_4) of the surface along incongruence line shows how the degree of discrepancy between agency and communion may influence reporting behavior. A significant negative curvature suggests that reporting behavior decreases as the strengths of agency and communion diverge and increases as their strengths converge (moving closer to the congruence line). The slope (a_3) of the surface along incongruence line indicates how the direction of the discrepancy influences reporting behavior. Third, the first principal axis indicates the ridge of a concave surface where the downward curvature is the least. If the intercept (p_{10}) of the first principal axis equals zero and the slope (p_{11}) equals one, it indicates that a ridge with the first principal axis along the congruence line. Otherwise, it indicates a translation or rotation of the ridge relative to the congruence line.

Main analysis results

Before testing the research model, considering the variables vary in their magnitudes, we standardized the independent variables for the main analysis. Meanwhile, in order to display the levels of the agency and communion more clearly, we report the descriptive statistical results in Table 2 before centring and the independent variables. Table 2 presents the descriptive statistics of our key variables and the correlation matrix.

We employed hierarchical regression to validate the appropriateness of polynomial regression for our study. The analysis results from Table 3 indicate that the variance of the quadratic model (i.e.,

Model 3; R^2 = 0.1106) exceeds that of the linear model (i.e., Model 2; R^2 = 0.0719), suggesting it is appropriate to perform polynomial regression for our study.

Hypothesis 1 proposed that the incongruence between agency and communion is negatively related to fake news reporting behavior. The results from Model 3 in Table 3 indicate that the curvature (a_4 = -0.0090, p < .01) along the incongruence line is significantly negative, as illustrated in Figure 1, where the response surface curves downward along the incongruence line. The results indicate that users' fake news reporting behavior is reduced when the strengths of agency and communion become further apart and is maximized when their strengths become more similar. Thus, hypothesis 1 is supported.

Variables	Mean	SD	1	2	3	4	5	6	7	8
1. Ln (ReportTimes+1)	0.7393	0.2335	-							
2. Agency	0.0005	0.0007	0.209***	-						
3. Communion	0.0004	0.0008	0.218***	0.302***	-					
4. Gender	0.6449	0.4785	-0.002	0.0003	-0.0002	-				
5. Ln (DaysSinceRegist)	8.1754	0.3694	-0.018**	0.0077	0.017**	-0.012	-			
6. Verification	0.0544	0.2268	0.062***	0.107***	0.108***	-0.026***	0.065***	-		
7. VIP	0.5544	1.7807	0.041***	0.047***	0.050***	-0.073***	0.049***	0.287***	-	
8. Credit	3.1438	0.4771	0.001	0.043***	0.027***	-0.044***	0.422***	0.263***	0.192***	-
9. Ln (NumberOfPosts)	5.7082	2.3292	0.025***	0.093***	0.072***	-0.155***	0.339***	0.243***	0.232***	0.391***

Note. The sample size of each variable in the table is 15256; *p < 10 percent, **p < 5 percent, ***p < 1 percent.

Table 2. Descriptive statistics and correlation matrix.

Hypothesis 2 proposed that users report more fake news when agency and communion congruence occur at higher level than lower level. The results from model 3 in Table 3 display a significantly positive slope (a_1 = 0.0956, p < .01) along the congruence line, suggesting that when agency and communion increase together at the same level, users will report more fake news.

Variables	Dependent variable: Ln (ReportTimes+1)				
	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	0.8685***	0.8593***	0.8686***	0.8493***	1.4085***
Gender	0.0015	-0.0005	-0.0018	-0.0025	-0.0001
Ln (DaysSinceRegist)	-0.0149**	-0.0126**	-0.0131**	-0.0104**	-0.0802*
Verification	0.0583***	0.0275***	0.0152*	-	-
VIP	0.0032***	0.0028**	0.0027**	0.0025**	0.0037
Credit	-0.0079*	-0.0059	-0.0054	-0.0085**	0.0508*
Ln (NumberOfPosts)	0.0020**	0.0001	-0.0009	-0.0002	-0.0142**
Agency		0.0362***	0.0434***	0.0308***	0.2031***
Communion		0.0392***	0.0521***	0.0539***	0.0744**
Agency × Agency			-0.0012***	-0.0007***	-0.0365***
Agency × Communion			0.0070***	0.0083***	0.0411***
Communion × Communion			-0.0008***	-0.0007***	-0.0085**
R^2	0.0055***	0.0719***	0.1106***	0.1053***	0.2011***
Sample size	15256	15256	15256	14426	830
Congruence line (Agency = Communion)					
Slope (a_1)			0.0956***	0.0848***	0.2776***
Curvature (a_2)			0.0049***	0.0068***	-0.0038
Incongruence line (Agency = -Communion)					
Slope (a_3)			-0.0086*	-0.0231***	0.1287***
Curvature (a_4)			-0.0090***	-0.0099***	-0.0862***
First principal axis					
Intercept (p_{10})			0.6952	2.3113*	-3.2596**
Slope (p_{11})			1.0543	1.0004	1.8874

Note. *p < 10 percent, **p < 5 percent, ***p < 1 percent.

Table 3. Polynomial regression results.

From Figure 1, it can be further observed that users' fake news reporting behavior exhibits a nonlinear increase along the congruence line from the front corner to the back corner on the response surface. This indicates that when users' agency and communion strengths are at the same level, any further increase will facilitate their reporting behavior of fake news. Thus, hypothesis 2 is supported.

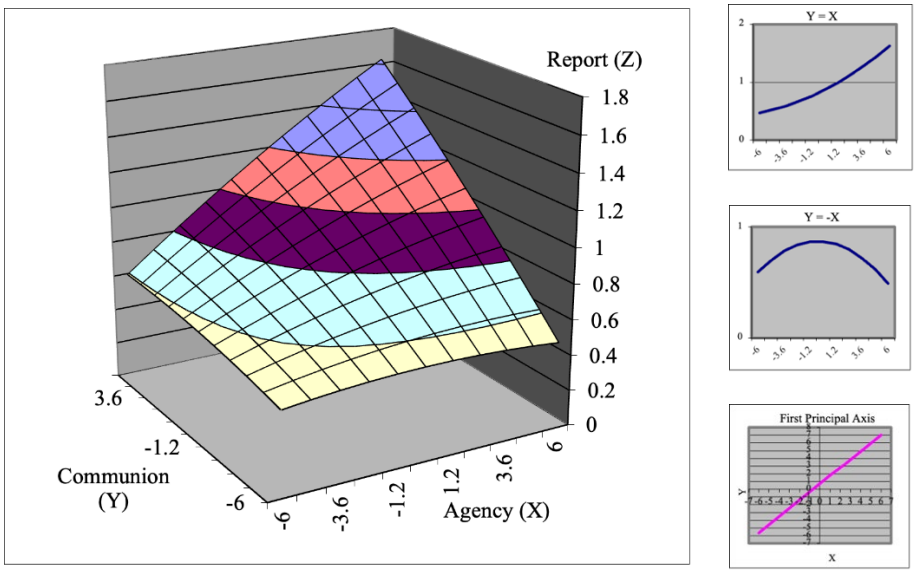


Figure 1. Response surface of agency and communion to reporting behavior (Model 3).

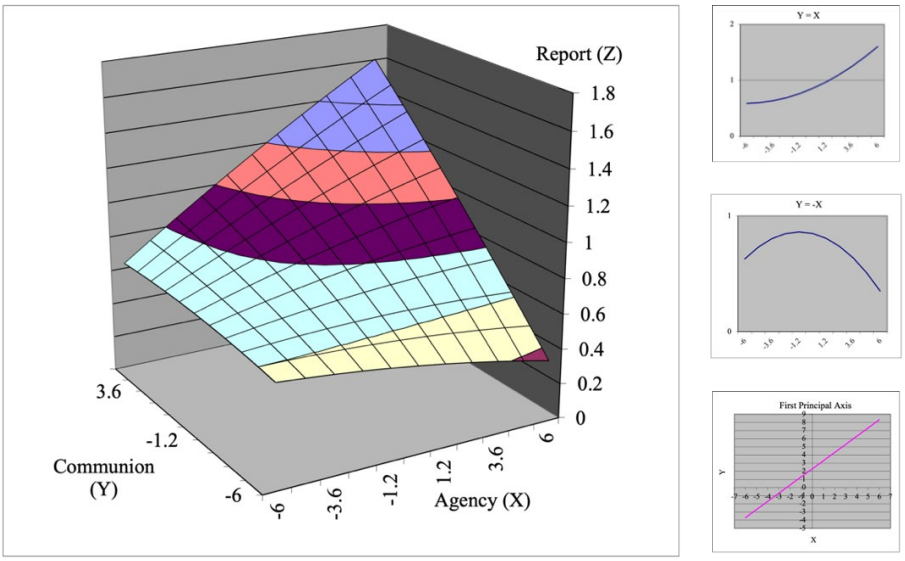


Figure 2. Response surface of agency and communion to reporting behavior (Model 4).

Hypothesis 3a proposed that unverified reporters report more fake news when their communion exceeds agency. The results from model 4 in Table 3 display a significantly negative slope (a_3 = -0.0231, p < .01) and curvature (a_4 = -0.0099, p < .01) along the incongruence line, as manifested in Figure 2, showing reporting behavior initially rises slowly and then declines rapidly along the line of incongruence from the back corner to the front corner on the surface. Because the curvature along incongruence line is significantly negative, the surface has a ridge. Thus, we further examine the first principal axis to see whether it is coincident with the congruence line. The purpose is to check the region where more reporting behavior occurs. As shown in Table 3, the first principal axis is $y = -1.0004x + 2.3113$. Its slope (p_{11} = -1.0004) is not significantly different from 1, but the intercept (p_{10} = 2.3113) is significantly different from 0, indicating that the ridge of the response surface does not run along the congruence line but instead parallel move towards the region where communion exceeds agency. The above results collectively corroborate that when communion exceeds agency, unverified users are more inclined to engage in reporting behaviors. Thus, Hypothesis 3a is supported.

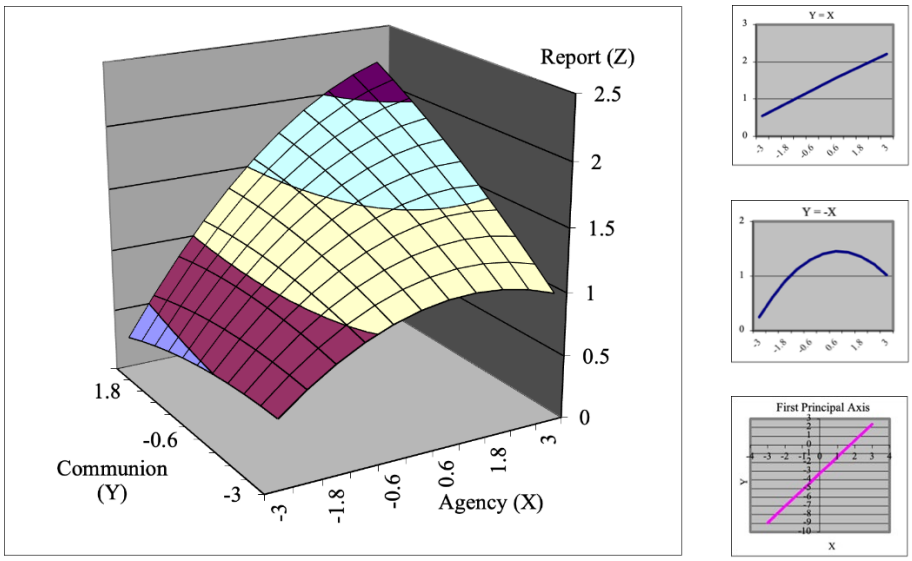


Figure 3. Response surface of agency and communion to reporting behavior (Model 5).

We tested hypothesis 3b using the same procedure. The results from model 5 in Table 3 display a significantly positive slope (a_3 = 0.1287, p < .01) and negative curvature (a_4 = -0.0862, p < .01) along the incongruence line, as manifested in Figure 3, showing reporting behavior initially rises rapidly and then declines slowly along the line of incongruence from the back corner to the front corner on the surface. To check the region where more reporting behavior occurs, we also examined the first principal axis. As shown in Table 3, the first principal axis is $y = -3.2596 + 1.8874x$, whose slope (p_{11} = 1.8874) is not significantly different from 1 and the intercept (p_{10} = -3.2596) is significantly different from 0. This represents that the ridge of the surface does not run along the congruence line but instead parallel move towards the region where agency exceeds communion. The above results corroborate that verified users will report more fake news when moral motivation of agency exceeds communion. Thus, Hypothesis 3b is supported.

Discussion

Theoretical implications

This study presents four key theoretical implications. First, it offers a novel perspective on fake news reporting behavior through the lens of moral motivation. While previous studies have focused on the antecedents of online reporting behavior at the social, platform, or content levels, such as educational interventions, interface design, or online content formats, few have considered moral motivation. Given that online reporting is a prosocial behavior reflecting concerns with social norms and moral responsibility, this study highlights the role of moral motivation. Using fake news reporting as the research context, this study explores how moral motivation influences users' reporting behaviors, providing new insights and contributing to existing research in this field.

Second, this study clarifies the debate on how the agency and communion dimensions of moral motivation influence moral behavior. Drawing on the reconciliation model of moral centrality, this study offers a comprehensive analysis of how the congruence or incongruence between agency and communion influences fake news reporting behavior. Based on unique field data, including over 13,000 reports and more than 4 million posts from 15,256 reporters, the findings show that incongruence between agency and communion negatively impacts fake news reporting behavior. Conversely, when agency and communion motivations are congruent, the likelihood of fake news reporting increases as the congruence between these dimensions grows. This study reveals the complex mechanisms through which the agency and communion dimensions of moral motivations influence fake news reporting behavior by demonstrating the existence of a curvilinear relationship.

Third, the analysis of user heterogeneity further reveals the differential impacts of the agency and communion on individuals' moral behavior. The results show that unverified social media users report more fake news when their moral motivation of communion is higher than agency. Conversely, verified social media users report more fake news when their moral motivation of agency outweighs communion. This finding illustrates the varying impact of moral motivations on distinct user groups, highlighting how user heterogeneity moderates the effects of agency and communion on fake news reporting behavior. It also contributes to ongoing debates on the agency and communion dimensions, showing their differentiated influence across different user categories.

Finally, this study is the first to employ PR-RSA to examine the impact of moral motivation duality on individual behavior. Unlike previous studies that calculated agency and communion as a single variable through subtraction or division, PR-RSA maintains the distinctiveness of these variables and reveal the complex curvilinear relationships between them and reporting behavior. In this regard, PR-RSA overcomes the limitations of traditional statistical models, providing a unique opportunity to understand the relationship between moral motivation and reporting behavior. We

encourage future research to adopt PR-RSA to explore complex variable relationships, especially when the congruence or incongruence of these variables has a significant impact.

Practical implications

First, this study shows that social media users whose agency and communion motivations are incongruent are less likely to engage in fake news reporting behavior. When the congruence between these two motivations is stronger, users are more likely to report fake news. Therefore, we suggest that social media platforms should emphasize the dual value of online reporting behaviors for both individuals and society, thus helping users better recognize that reporting fake news can serve both public and personal interests. Platforms could implement mechanisms, such as certification or reputation-building systems, to help users enhance their personal standing through the act of reporting fake news.

Second, the findings also suggest that the moral motivations influencing fake news reporting differ for verified and unverified social media users. Therefore, social media platforms can adopt tailored incentive strategies for each user group. For unverified users, platforms should emphasize the potential harm that fake news can cause to vulnerable groups, thus enhancing their moral motivation of communion. In contrast, for verified users, incentives focused on self-interest, such as access to exclusive content or other tangible rewards, should be used to bolster their moral motivation of agency.

Limitations and future research

This study has two limitations, which present opportunities for future research. First, we only used data collected from Chinese social media platforms, and the relatively homogeneous social context may raise concerns about the generalizability of the research results. Therefore, we recommend that future research attempt to use data from other cultural backgrounds or countries to test the validity of this study. Second, the focus of this study was to explain whether the users' agency and communion motivations are fully reconciled to influence the fake news reporting behavior. However, it did not empirically test which strategies could promote the integration of agency and communion among social media users. Future researchers might continue to explore how to effectively promote the full integration of these motivations among users.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China [72274144, 72311540158], the Humanities and Social Sciences Foundation of the Ministry of Education, China [22YJA870013], and the Fundamental Research Funds for the Central Universities. We also greatly appreciate Dr. Feng Yu and Dr. Liang Zhao for providing the latest Chinese version of the Moral Motivation Dictionary, which was used to calculate the agency and communion dimensions of moral motivation.

About the authors

Xiao-Liang Shen is a professor in the School of Information Management at Wuhan University. His current research interests include information management, the dark side of IT, and digital governance. He has published in journals such as Journal of the Association for Information Systems, Journal of Information Technology, Journal of the Association for Information Science and Technology, Information & Management, and Decision Support Systems. He can be contacted at xlshen@whu.edu.cn

Lin-Yao Liu is a Ph.D. student at the School of Information Management, Wuhan University. She received her master's degree in 2022. Her current research focuses on information management,

behavior research, and the study of disinformation and misinformation on social media. She can be contacted at lyliu1010@whu.edu.cn

Yang-Jun Li is an assistant professor in the School of Management at Beijing Institute of Technology. His research interests include ethical issues in the use and application of IT, e-commerce, and social media. He has published in journals such as Journal of the Association for Information Systems, Journal of the Association for Information Science and Technology, and Information & Management. He can be contacted at liyangjun@bit.edu.cn

References

Bauer, J. J., & Wayment, H. A. (2008). The psychology of the quiet ego. In H. A. Wayment & J. J. Bauer (Eds.), Transcending self-interest: Psychological explorations of the quiet ego. (pp. 7–19). American Psychological Association. https://doi.org/10.1037/11771-001

Bell, E. (2024). Against Moral Individualism: Special Relations and the Agent-Neutral/Agent-Relative Distinction. Social Theory & Practice, 50(1). 10.5840/soctheorpract20231122206

Bellah, R. N., Madsen, R., Sullivan, W. M., Swidler, A., & Tipton, S. M. (2007). Habits of the heart, with a new preface: Individualism and commitment in American life. Univ of California Press.

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. Academy of Management Journal, 36(6), 1577–1613. https://doi.org/10.2307/256822

Flanagan, O. (2009). Varieties of moral personality: Ethics and psychological realism. Harvard University Press.

Frimer, J. A., Schaefer, N. K., & Oakes, H. (2014). Moral actor, selfish agent. Journal of Personality and Social Psychology, 106(5), 790–802. https://doi.org/10.1037/a0036040

Frimer, J. A., & Walker, L. J. (2009). Reconciling the self and morality: An empirical model of moral centrality development. Developmental Psychology, 45(6), 1669–1681. https://doi.org/10.1037/a0017418

Frimer, J. A., Walker, L. J., Dunlop, W. L., Lee, B. H., & Riches, A. (2011). The integration of agency and communion in moral personality: Evidence of enlightened self-interest. Journal of Personality and Social Psychology, 101(1), 149. https://doi.org/10.1037/a0023780

Frimer, J. A., Walker, L. J., Lee, B. H., Riches, A., & Dunlop, W. L. (2012). Hierarchical Integration of Agency and Communion: A Study of Influential Moral Figures. Journal of Personality, 80(4), 1117–1145. https://doi.org/10.1111/j.1467-6494.2012.00764.x

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The Effectiveness of Social Norms in Fighting Fake News on Social Media. Journal of Management Information Systems, 38(1), 196–221. https://doi.org/10.1080/07421222.2021.1870389

Gupta, P., Ding, B., Guan, C., & Ding, D. (2024). Generative AI: A Systematic Review using Topic Modelling Techniques. Data and Information Management, 8(2), 100066. https://doi.org/10.1016/j.dim.2024.100066

Jiang, T., Sun, Z., Fu, S., & Lv, Y. (2024). Human-AI Interaction Research Agenda: A User-Centered Perspective. Data and Information Management, 8(4), 100078. https://doi.org/10.1016/j.dim.2024.100078

Kankanamge, K. S. I., Chowdhury, A., Kabir, K. H., & Khan, N. A. (2025). Evils of knowledge sharing and learning: The case of agri-food misinformation in virtual communities of practices in Sri Lanka. Data and Information Management, 100090. https://doi.org/10.1016/j.dim.2024.100090

Kim, A., Dennis, A. R., & Indiana University. (2019). Says Who? The Effects of Presentation Format and Source Rating on Fake News in Social Media. MIS Quarterly, 43(3), 1025–1039. https://doi.org/10.25300/MISQ/2019/15188

Kim, J. Y., Sim, J., & Cho, D. (2023). Identity and Status: When Counterspeech Increases Hate Speech Reporting and Why. Information Systems Frontiers, 25(5), 1683–1694. https://doi.org/10.1007/s10796-021-10229-2

Lapsley, D. K. (2004). Moral functioning: Moral understanding and personality. In Moral development, self, and identity (pp. 347–360). Psychology Press.

Li, Y. J., Cheung, C. M., Shen, X. L., & Lee, M. K. (2024). Doxing on Social Networking Sites: An Extension of the Social Cognitive Theory of Moral Thought and Action. Journal of the Association for Information Systems, 25(6), 1466–1499. https://doi.org/10.17705/1jais.00898

Li, Y. J., Marga, J. J., Cheung, C. M., Shen, X. L., & Lee, M. K. (2022). Health Misinformation on Social Media: A Systematic Literature Review and Future Research Directions. AIS Transactions on Human-Computer Interaction, 14(2), 116-149. https://doi.org/10.17705/1thci.00164

Liang, H., Wang, N., & Xue, Y. (2022). Juggling Information Technology (IT) Exploration and Exploitation: A Proportional Balance View of IT Ambidexterity. Information Systems Research, 33(4), 1386–1402. https://doi.org/10.1287/isre.2022.1105

Liu, L., Mirkovski, K., Lowry, P. B., & Vu, Q. (2023). "Do as I say but not as I do": Influence of political leaders' populist communication styles on public adherence in a crisis using the global case of COVID-19 movement restrictions. Data and Information Management, 7(2), 100039. https://doi.org/10.1016/j.dim.2023.100039

Naderer, B., Wendt, R., Bachl, M., & Rieger, D. (2023). Understanding the role of participatory-moral abilities, motivation, and behavior in European adolescents' responses to online hate. New Media & Society, 14614448231203617. https://doi.org/10.1177/14614448231203617

Rodriguez Müller, A. P., Lerusse, A., Steen, T., & Van De Walle, S. (2024). What motivates users to report service-related issues? A study on coproduction in a smart public service. Public Administration, 102(2), 691–714. https://doi.org/10.1111/padm.12946

Sanders, S., Wisse, B., Van Yperen, N. W., & Rus, D. (2018). On Ethically Solvent Leaders: The Roles of Pride and Moral Identity in Predicting Leader Ethical Behavior. Journal of Business Ethics, 150(3), 631–645. https://doi.org/10.1007/s10551-016-3180-0

Shen, X. L., & Wu, Y. (2024). From whispers to warriors: Public atmosphere's role in mobilizing social media users against rumors. Telematics and Informatics, 90, 102133. https://doi.org/10.1016/j.tele.2024.102133

Sun, J., Wilt, J., Meindl, P., Watkins, H. M., & Goodwin, G. P. (2024). How and why people want to be more moral. Journal of Personality, 92(3), 907–925. https://doi.org/10.1111/jopy.12812

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, 29(1), 24–54. https://doi.org/10.1177/0261927X09351676

Taylor, C. (1989). Sources of the self: The making of the modern identity. Harvard University Press.

Walker, L. J. (2013). Exemplars' Moral Behavior Is Self-Regarding. New Directions for Child and Adolescent Development, 2013(142), 27–40. https://doi.org/10.1002/cad.20047

Walker, L. J., & Frimer, J. A. (2007). Moral personality of brave and caring exemplars. Journal of Personality and Social Psychology, 93(5), 845–860. https://doi.org/10.1037/0022-3514.93.5.845

Wang, S. (Ada), Pang, M.-S., & Pavlou, P. A. (2021). Cure or Poison? Identity Verification and the Posting of Fake News on Social Media. Journal of Management Information Systems, 38(4), 1011–1038. https://doi.org/10.1080/07421222.2021.1990615

Wang, S. (Ada), Pang, M.-S., Temple University, Pavlou, P., & University of Houston. (2022). Seeing Is Believing? How Including a Video in Fake News Influences Users' Reporting of Fake News to Social Media Platforms. MIS Quarterly, 45(3), 1323–1354. https://doi.org/10.25300/MISQ/2022/16296

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting Hate Comments: Investigating the Effects of Deviance Characteristics, Neutralization Strategies, and Users' Moral Orientation. Communication Research, 47(6), 921–944. https://doi.org/10.1177/0093650219855330

Wong, R. Y. M., Cheung, C. M. K., Xiao, B., & Thatcher, J. B. (2021). Standing Up or Standing By: Understanding Bystanders' Proactive Reporting Responses to Social Media Harassment. Information Systems Research, 32(2), 561–581. https://doi.org/10.1287/isre.2020.0983

Wu, S. (2024). What motivates audiences to report fake news?: Uncovering a framework of factors that drive the community reporting of fake news on social media. Digital Journalism, 12(6), 790–807. https://doi.org/10.1080/21670811.2023.2243489

Wu, Y., Shen, X. L., & Sun, Y. (2023). Establishing the typology and the underlying structure of rumor-combating behaviors: A multidimensional scaling approach. Information Technology & People, 36(7), 2661–2686. https://doi.org/10.1108/ITP-01-2022-0072

Zhang, Y., & Yu, F. (2018). Which Socio-Economic Indicators Influence Collective Morality? Big Data Analysis on Online Chinese Social Media. Emerging Markets Finance and Trade, 54(4), 792–800. https://doi.org/10.1080/1540496X.2017.1321984

Zhao, L., Ding, X., & Yu, F. (2020). Public moral motivation during the COVID-19 pandemic: Analysis of posts on Chinese social media. Social Behavior and Personality: An International Journal, 48(11), 1–14. https://doi.org/10.2224/sbp.9829

Zhou, H., Lu, Y., Zhao, L., Wang, B., & Li, T. (2023). Effective reporting system to encourage users' reporting behavior in social media platforms: An empirical study based on structural empowerment theory. Behaviour & Information Technology, 43(14), 3490-3509. https://doi.org/10.1080/0144929X.2023.2281491

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Drivers of metaverse adoption among children and youths: a UTATU2 perspective

Liuyu Huang, Dion Hoe-Lian Goh, and Stella Xin Yin
DOI: https://doi.org/10.47989/ir30iConf47593

Abstract

Introduction. Despite increasing commercial and research interest in the metaverse, its adoption has been limited due to challenges such as technological complexities, usability issues, cost, and privacy concerns. This study investigates factors influencing metaverse adoption from the perspectives of children and youths in Asian countries.

Method. Using the *unified theory of acceptance and use of technology* 2 (UTAUT2), we conducted semi-structured interviews with 40 individuals aged 13 to 24 to understand their metaverse usage habits and the perceived drivers of adoption.

Analysis. Audio recordings of all interviews were transcribed. The transcripts were manually content analysed through an iterative process. Themes were based on the UTAUT2.

Results. Entertainment, socialization, and education were primary activities conducted within the metaverse. Adoption was mainly driven by performance expectancy, facilitating conditions, and social factors.

Introduction

The term '*metaverse*' has gained significant attention in recent years. For instance, the governments of South Korea and Dubai have invested millions in developing metaverse ecosystems (Piscione & Drean, 2023). Scholars also predict that the metaverse trend will continue to grow (Sami et al., 2024). However, user adoption has not progressed as quickly as expected. Metaverse adoption in educational settings, for example, remains low (Talan & Kalinkara, 2022), largely due to challenges related to technology, usability, cost, and privacy concerns (Al-Kfairy et al., 2024; Iwanaga et al., 2023).

As a result, there has been a growing body of research focused on understanding the drivers of metaverse adoption. Drivers are factors that encourage, facilitate, or support individuals in their adoption of technologies (Makransky & Mayer, 2022). Examples include perceived enjoyment and satisfaction (Makransky & Mayer, 2022), as well as user expectations of performance (Lee, Trimi, & Kim, 2013). Such research is crucial because it offers strategies to enhance the adoption of potentially transformative technologies and provides valuable insights into ensuring their sustained usage.

There are three research gaps in extant work on the drivers of metaverse adoption. First, existing studies predominantly focused on general users or university students (Alawadhi et al, 2023; Almarzouqi et al, 2022) with limited exploration of younger age groups. To date, only a few empirical studies specifically addressed adolescents' perceptions of the metaverse: two studies examined the effectiveness of metaverse applications in educational settings (Makransky & Mayer, 2022; George Reyes, 2020), one proposed a prototype for a metaverse-based physical fitness service for youth (Lee et al， 2023), another investigated why teenagers remained engaged in social virtual reality (Maloney et al, 2021), and the final one reviewed the benefits and potential concerns among adolescents (Kim & Kim, 2023).

Second, most empirical research has been conducted in Europe (Makransky & Mayer, 2022), North America (George Reyes, 2020; Maloney et al, 2021), and Gulf-region countries (Alawadhi et al., 2023; Almarzouqi et al, 2022), with minimal representation from the Asian region. Given that cultural differences influence technology adoption (Sewandono et al., 2023), studying metaverse adoption in Asia could significantly enrich the existing literature. Notably, only one study was Asian-focused, examining metaverse acceptance among Korean university students (Lee & Kim, 2022). Third, existing research on children's and youths' adoption of the metaverse has been largely confined to educational settings, with few studies investigating usage outside of this context.

To address these gaps, the present study addresses two research questions: (1) what activities do children and youths engage in the metaverse? (2) what factors drive their adoption? To answer these questions, we employ the *unified theory of acceptance and use of technology* 2 (UTAUT2) as our theoretical framework (Venkatesh, Thong, & Xu, 2012). The UTAUT2 extends the original *UTAUT* model by incorporating three new factors - hedonic motivation, price value, and habit - thus enhancing its applicability to both organizational and non-organizational settings (Venkatesh, Thong & Xu, 2012; Blok et al., 2020).

Literature review

Metaverse concepts

The metaverse is an immersive, 3D virtual world where users, represented by avatars, interact with online media, applications, and each other in real-time, creating a personal sense of presence (Abbate et al., 2022). This concept is characterized by three key features: immersion, interaction, and presence. As the anticipated next iteration in the evolution of the internet (Arpaci et al., 2022), metaverse applications have been employed across various industries, including healthcare, education, and engineering (Ning et al., 2023).

Existing research on metaverse usage among children and youths has primarily concentrated on psychological and physiological issues. Studies have examined both the benefits and risks, such as a prototype of metaverse-based training system for youth physical fitness management (Lee et al, 2023), concerns about harassment and privacy (Maloney et al, 2021), and issues related to metaverse addiction, antisocial behavior like cyberbullying, and identity confusion (Kim & Kim, 2023). In contrast, the present study shifts the focus to uncovering the activities children and youths engage in with the metaverse, as well as the factors driving their adoption. Our findings provide a foundation for a better understanding of how these age groups can more effectively engage with the metaverse.

Unified theory of acceptance and use of technology 2

The UTAUT2 is one of the most widely used frameworks for studying technology adoption. Building on the original UTAUT, which included factors such as performance expectancy, effort expectancy, social influence, and facilitating conditions, the UTAUT2 introduces three additional factors: hedonic motivation, price value, and habit (Venkatesh, Thong, & Xu, 2012). The UTAUT2 was selected because of its consistency and robustness in explaining usage intention and actual use of new technology when compared to other frameworks (Blok et al., 2020; Abbate et al., 2022).

The UTAUT2 has been applied across various technology contexts, including smartwatches (Beh et al., 2021), smart homes (Nascimento et al., 2023), and blockchain-based remittance services (Mansoor et al., 2023). In the context of metaverse-related applications, it has been used to study adoption within the education sector. For instance, Sunardi et al. (2022) investigated students' adoption of augmented reality in a video conference learning course during the COVID-19 pandemic, while Teng et al. (2023) examined the adoption of mixed reality systems by students enrolled in medical courses.

Methodology

Qualitative interviews were conducted with 20 children between 13 and 17 years old (represented by 'C' and a unique number in subsequent sections), and 20 youths between 18 and 24 years old (represented by 'Y' and a unique number) to understand their metaverse usage habits and perceived drivers of metaverse adoption. The interview method was chosen because it allows for an in-depth examination of participants' experiences, considerations, and interpretations, thereby enhancing our understanding of metaverse use and perceptions (Lee, Kang, & Kim, 2023). Participants were recruited from the Asia-Pacific region including Singapore, India, and China.

Participants consisted of 19 females and 21 males, with 38 still in school, ranging from secondary school to university levels, and two who had just entered the workforce. Parental consent was obtained for participants under 21 years old, while those 21 and above provided their own consent. Recruitment was carried out using snowball sampling, starting with the researchers' network, and expanding through voluntary referrals from participants at the end of each interview.

All interviews were conducted online via Zoom, each lasting approximately 45 minutes. The audio recordings were initially transcribed using Zoom's caption function and then manually corrected by a researcher. The transcripts were analyzed through an iterative process to ensure internal consistency. A researcher and a trained student assistant jointly coded four transcripts (10% of the total) based on themes related to UTUAT2. During this process, disagreements were highlighted, discussed, and resolved. Subsequently, the remainder of the transcripts were equally divided between the two coders for analysis. The intercoder reliability achieved a 93% agreement. This approach aligns with existing content analysis practices (Stemler, 2000).

Conclusions

Metaverse usage patterns

Interestingly, games served as the primary gateway to the metaverse for many participants. Specifically, 28 of our 40 participants reported engaging with metaverse platforms primarily through free-to-play, open-world, action role-playing game platforms. The most frequently mentioned games were Roblox, Minecraft, Genshin impact, and Mobile legends. Participants typically played on mobile devices or desktop computers. Notably, only one participant reported using VR headsets to play tennis for skills and rules practice. Participants generally used metaverse platforms weekly, spending 2-3 hours per session.

Unsurprisingly, entertainment was the most common reported reason for using the metaverse. This category included two aspects: playing role-playing games with friends, such as killing creatures, collecting items, or completing missions, as well as creating new experiences. For instance, participant C3 (male, 13 years old) noted, '*metaverse games are very realistic, and it's fun - you can play with your friends, or you can play with anyone on the internet.*' Additionally, two participants mentioned creating their own games on Roblox: C12 (male, 14 years old) built a house in Roblox and invited his friends to decorate it, while C4 (male, 13 years old) developed a game for physics simulation.

Socializing was the second most reported reason for metaverse use, encompassing interactions with friends, family, or even strangers. Games often facilitated these social connections, as noted by C15 (female, 16 years old), who played metaverse games to engage with friends who enjoyed gaming. Chat functionality was also a popular method for socializing, with nine participants reporting chatting with random people online. Education was the third most common but reported by only four participants who used the metaverse as part of school or training requirements. Finally, two participants explored virtual environments with no fixed aim to pass the time.

Metaverse drivers

Within the UTAUT2 framework of metaverse adoption (Blok et al., 2020), the primary factors identified from the interviews were performance expectancy, social relations and social influence, and facilitating conditions. In contrast, effort expectancy, perceived value, and habit were less frequently mentioned.

Performance expectancy

Performance expectancy refers to the extent to which users believe that using technology will help to improve job performance goals (Lee et al, 2013). Put simply, it refers to the usefulness of a technology. Participants highlighted three main themes under performance expectancy as drivers for metaverse adoption. First, the metaverse was reported as a means of stress relief. This was the most frequently mentioned reason for using metaverse platforms, with participants indicating that the metaverse offered a valuable break or a nice distraction from work or school. As one participant put it succinctly, playing games in the metaverse is '*just like a form of relieving stress*' (Y11, female, 20 years old).

The second-most common theme was the metaverse's role as a tool for building and maintaining relationships. Participants utilized the metaverse to connect with friends, family, or unknown people globally. It was described as '*a little bonding session with my friends*' (Y13, male, 20 years old) and as '*an icebreaker or a warm-up activity for classmates to get to know each other*' (Y4, female, 20 years old). Lastly, the metaverse was acknowledged as offering new ways to gain knowledge. For instance, C13 (male, 13 years old) noted that a game could simulate real-world scenarios like managing finance, where users must budget and control their spending.

Social relations and social influence

Social relations refer to the relationships or interactions between individuals, while social influence refers to the way individuals are affected by those, they spend time with (Montgomery et al., 2020). Participants' close relationships with their friends, family members, and peers played a significant role in their adoption of the metaverse. Many participants reported being introduced to the metaverse by friends or family. For example, Y7, a 20-year-old male, began exploring the metaverse via World of Warcraft due to his friends' interest in such platforms. To join them, he also started playing similar games, a trend echoed by four other children and two youths.

Facilitating conditions

Facilitating conditions involve users' perceptions of the availability of the resources and support necessary for engaging with technology (Venkatesh et al, 2012). In this study, this concept encompasses the personnel and educational resources available to assist with metaverse use. Personnel includes teachers, coaches, and support staff who aid in using metaverse hardware and software, while educational resources consist of demonstrations, tutorials, and other material to assist users. Given the unique and novel nature of virtual reality experiences offered by metaverse platforms (Montgomery et al., 2020), users often require support, particularly during the initial periods. For instance, Y3 (male, 24 years old) received training on using a VR headset and remote controller and reported that coaches and support personnel were present to guide him and address equipment issues.

Other drivers

Three categories of drivers were less represented in the participants' responses: effort expectancy, perceived value, and habit. First, effort expectancy refers to the ease of using the technology (Venkatesh et al, 2012). For instance, Y10 (male, 20 years old) used Roblox and emphasized that the barriers to entry were '*not too high*'.

Next, perceived value involves the cognitive trade-offs between the benefits of the technology and its monetary cost (Venkatesh et al, 2012). Y11 (female, 20 years old), for example, spent money on avatar skins and decorations in Roblox, which he deemed '*valuable*', leading him to continue using and '*safeguarding*' them. Finally, habit is defined as the extent to which individuals automatically engage in certain behaviors (Venkatesh et al, 2012). Two participants reported that their enjoyment of metaverse games led to the formation of a habit of trying out new games, such as Villain, League of legends, and Mobile legends.

Discussion and conclusion

This study investigated the types of activities children and youths performed on metaverse platforms and the factors driving their adoption. Our findings revealed that the most frequently cited activities were entertainment, followed by socialization, education, and exploration. Using the UTAUT2 framework, we identified three major drivers for metaverse adoption: performance expectancy, social relations and social influence, and facilitating conditions. Effort expectancy, perceived value, and habit were mentioned less frequently by participants.

Given the dominance of gaming in metaverse usage, it was unsurprising that hedonic goals, namely, entertainment, socializing, and exploration were frequently cited reasons for engagement. Closer examination of participants' responses indicated that these motivations were often interconnected. For instance, in multi-user metaverse games, participants reported playing them for enjoyment and staying in touch with their friends, highlighting the appeal of such games in fulfilling multiple gratifications simultaneously (Yu, 2024).

Consistent with the UTATU2 literature (Abbad, 2021; Sewandono et al., 2023), performance expectancy emerged as the primary driver for metaverse adoption. This may be attributed to the metaverse's role as a technological tool for enhancing well-being and providing a social space for

meeting and making friends. Social influence was also found to be a significant driver of adoption (Okumus et al., 2018), potentially due to younger users being more sensitive to peer influence (Ciranka & Van den Bos, 2019) and experiencing a '*fear of losing out*' when others used the metaverse (Lee, Nguyen, & Yang, 2023). Moreover, in Asian countries where collectivistic cultures are predominant, people tend to rely on other likeminded individuals who have already adopted the innovation for their own decision making (Lee et al, 2013). Next, facilitating conditions, such as personnel and educational resources, were found to support metaverse adoption, aligning with recent studies that emphasized the importance of perceived support and resource accessibility in technology acceptance (Abbad, 2021; Patil et al, 2020).

It was interesting that effort expectancy, perceived value, and habit were the least mentioned drivers among participants. This suggests a lower emphasis on these factors compared to prior research (Yang et al, 2022). One possible explanation is that younger individuals are generally more open to new technology and are technologically savvy (Szymkowiak et al, 2021). To them, their focus is more on using new technology than about usability or value. Additionally, since many participants were school going, they likely did not have sufficient time to form habitual metaverse usage (Bayram, 2022). As well, examining habitual use of the metaverse may be more appropriate and productive during its growth or maturity stage. This is because with more people adopting a mature technology, there will likely be more opportunities to study habit formation and prolonged use (Tamilmani et al, 2019).

This study makes both theoretical and practical contributions. Theoretically, it extends the applicability of the UTAUT2 model to the metaverse context by adapting its factors to suit metaverse-specific use cases. It also enriches the metaverse adoption literature by including perspectives from children and youths in the Asia-Pacific region. Practically, the study suggests that metaverse companies focus on performance expectancy when developing platform functionality. Additionally, it recommends providing increased support personnel and training resources during the initial deployment of metaverse applications. This could be achieved through partnerships and collaborations between metaverse companies and educational institutions, for example, building an education-oriented metaverse ecosystem where teaching and learning occur virtually with interactive and shared open resources (Zhai et al., 2023). However, this study has limitations. The sample size was small, and the results may not be generalizable. Future research could address this issue by recruiting larger samples of participants and conducting quantitative surveys to validate the findings. Furthermore, as the study was conducted at a single point in time, future longitudinal research is needed to examine how perception of drivers evolves as children and youths mature in both age and metaverse experience.

Acknowledgements

This study is sponsored by Meta Platforms Inc with Grant No. INB2681518.

About the authors

Liuyu Huang is a research staff at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. Her research interests include bibliometrics, information studies. She can be contacted at lyhuang@ntu.edu.sg

Dion H. Goh is a Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. His research interests include social media practices and perceptions, game-based techniques for shaping user perceptions and motivating behaviour, and online information sharing and seeking. He can be contacted at ashlgoh@ntu.edu.sg

Stella Xin Yin is a PhD candidate at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. Her research interests include computational thinking, collaborative learning, and AI literacy. She can be contacted at xin013@e.ntu.edu.sg

References

Abbad, M. M. M. (2021). Using the UTAUT model to understand students' usage of e-learning systems in developing countries. Education and Information Technologies, 26(6), 7205–7224. https://doi.org/10.1007/s10639-021-10573-5.

Abbate, S., Centobelli, P., Cerchione, R., Oropallo, E. and Riccio, E. (2022), "A first bibliometric literature review on metaverse", Proceedings of the 2022 IEEE Technology and Engineering Management Conference, Izmir, Turkey, pp. 254-260, doi: 10.1109/TEMSCONEUROPE54743. 2022.9802015.

Al-Kfairy, M., Alomari, A., Al-Bashayreh, M., Alfandi, O., & Tubishat, M. (2024). Unveiling the Metaverse: A survey of user perceptions and the impact of usability, social influence, and interoperability. Heliyon.

Alawadhi, M., Alhumaid, K., Almarzooqi, S., Aljasmi, S., Aburayya, A., Salloum, S. A., & Almesmari, W. (2023). Factors affecting medical students' acceptance of the Metaverse system in medical training in the United Arab Emirates. South-Eastern European Journal of Public Health, 1–14. https://doi.org/10.56801/seejph.vi.310.

Almarzouqi, A., Aburayya, A., & Salloum, S. A. (2022). Prediction of user's intention to use Metaverse system in medical education: A hybrid SEM-ML learning approach. IEEE Access, 10, 43421–43434. https://doi.org/10.1109/ACCESS.2022.3169285.

Arpaci, I., Karatas, K., Kusci, I., & Al-Emran, M. (2022). Understanding the social sustainability of the Metaverse by integrating UTAUT2 and big five personality traits: A hybrid SEM-ANN approach. Technology in Society, 71, 102120.

Bayram, A. (2022). Metaleisure: Leisure time habits to be changed with metaverse. Journal of Metaverse, 2(1), 1-7.

Beh, P. K., Ganesan, Y., Iranmanesh, M., & Foroughi, B. (2021). Using smartwatches for fitness and health monitoring: the UTAUT2 combined with threat appraisal as moderators. Behavior & Information Technology, 40(3), 282-299.

Blok, M., van Ingen, E., de Boer, A. H., & Slootman, M. (2020). The use of information and communication technologies by older people with cognitive impairments: from barriers to benefits. Computers in Human Behavior, 104, 106173.

Ciranka, S., & Van den Bos, W. (2019). Social influence in adolescent decision-making: A formal framework. Frontiers in Psychology, 10, 467793.

George Reyes, C. E. (2020). High school students' views on the use of metaverse in mathematics learning. Metaverse, 2(1), 9. https://doi.org/10.54517/m.v1i2.1777.

Hanif, M. S., Wang, M., Mumtaz, M. U., Ahmed, Z., & Zaki, W. (2022). What attracts me or prevents me from mobile shopping? An adapted UTAUT2 model empirical research on behavioral intentions of aspirant young consumers in Pakistan. Asia Pacific Journal of Marketing and Logistics, 34(5), 1031-1059.

Iwanaga, J., Muo, E. C., Tabira, Y., Watanabe, K., Tubbs, S. J., D'Antoni, A. V., ... & Tubbs, R. S. (2023). Who really needs a Metaverse in anatomy education? A review with preliminary survey results. Clinical Anatomy, 36(1), 77-82. https://doi.org/10.1002/ca.23949.

Kim, S., & Kim, E. (2023). Emergence of the Metaverse and psychiatric concerns in children and adolescents. Journal of the Korean Academy of Child and Adolescent Psychiatry, 34(4), 215.

Lee, J., Yoon, H. K., & Kim, D. (2023). Design of Metaverse-Based Physical Fitness Service for the Enhancement of Exercise Capability for Youth. Mobile Information Systems, 2023(1), 7272781.

Lee, S. G., Trimi, S., & Kim, C. (2013). The impact of cultural differences on technology adoption. Journal of World Business, 48(1), 20-29. https://doi.org/10.1016/j.jwb.2012.06.003.

Lee, S. H., Kang, I., & Kim, H. W. (2023). Understanding cybercrime from a criminal's perspective: Why and how suspects commit cyber-crimes? Technology in Society, 75, 102361.

Lee, U. K., & Kim, H. (2022). UTAUT in Metaverse: an 'Ifland' case. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 613-635. https://doi.org/10.3390/jtaer17020032.

Lee, Y. C., Nguyen, M. N., & Yang, Q. (2023). Factors Influencing Vietnamese Generation MZ's Adoption of Metaverse Platforms. Sustainability, 15(20), https://doi.org/10.3390/su152014940.

Makransky, G., & Mayer, R. E. (2022). Benefits of taking a virtual field trip in immersive virtual reality: Evidence for the immersion principle in multimedia learning. Educational Psychology Review, 34(3), 1771–1798. https://doi.org/10.1007/s10648-022-09675-4.

Maloney, D., Freeman, G., & Robb, A. (2021, June). Stay connected in an immersive world: Why teenagers engage in social virtual reality. In Proceedings of the 20th Annual ACM Interaction Design and Children Conference (pp. 69-79).

Mansoor, M., Abbasi, A. Z., Abbasi, G. A., Ahmad, S., & Hwang, Y. (2023). Exploring the determinants affecting the usage of blockchain-based remittance services: an empirical study on the banking sector. Behavior & Information Technology, 1-19.

Montgomery, S. C., Donnelly, M., Bhatnagar, P., Carlin, A., Kee, F., & Hunter, R. F. (2020). Peer social network processes and adolescent health behaviors: A systematic review. Preventive Medicine, 130, 105900.

Nascimento, D. R., Ciano, M. P., Gumz, J., & Fettermann, D. C. (2023). The acceptance process of smart homes by users: a statistical meta-analysis. Behavior & Information Technology, 42(16), 2768-2785.

Ning, H., Wang, H., Lin, Y., Wang, W., Dhelim, S., Farha, F., ... & Daneshmand, M. (2023). A Survey on the Metaverse: The State-of-the-Art, Technologies, Applications, and Challenges. IEEE Internet of Things Journal. https://doi.org/10.1109/jiot.2023.3278329.

Okumus, B., Ali, F., Bilgihan, A., & Ozturk, A. B. (2018). Psychological factors influencing customers' acceptance of smartphone diet apps when ordering food at restaurants. International Journal of Hospitality Management, 72, 67-77. https://doi.org/10.1016/j.ijhm.2018.01.001.

Patil, P., Tamilmani, K., Rana, N. P., & Raghavan, V. (2020). Understanding consumer adoption of mobile payment in India: Extending Meta-UTAUT model with personal innovativeness, anxiety, trust, and grievance redressal. International Journal of Information Management, 54, 102144.

Piscione, D. P., & Drean, J. (2023, May 12). Yes, the Metaverse Is Still Happening. Harvard Business Review. Retrieved August 12, 2024, from https://hbr.org/2023/05/yes-the-metaverse-is-still-happening.

Sami, H., Hammoud, A., Arafeh, M., Wazzeh, M., Arisdakessian, S., Chahoud, M., ... & Guizani, M. (2024). The metaverse: Survey, trends, novel pipeline ecosystem & future directions. IEEE Communications Surveys & Tutorials.

Schmitz, A., Díaz-Martín, A. M., & Guillén, M. J. Y. (2022). Modifying UTAUT2 for a cross-country comparison of telemedicine adoption. Computers in Human Behavior, 130, 107183.

Sewandono, R. E., Thoyib, A., Hadiwidjojo, D., & Rofiq, A. (2023). Performance expectancy of E-learning on higher institutions of education under uncertain conditions: Indonesia context. Education and information technologies, 28(4), 4041-4068.

Stemler, S. (2000). An overview of content analysis. Practical Assessment, Research, and Evaluation, 7(1), 17.

Sunardi, S., Ramadhan, A., Abdurachman, E., Trisetyarso, A., & Zarlis, M. (2022). Acceptance of augmented reality in video conference-based learning during COVID-19 pandemic in higher education. Bulletin of Electrical Engineering and Informatics, 11(6), 3598-3608.

Szymkowiak, A., Melović, B., Dabić, M., Jeganathan, K., & Kundi, G. S. (2021). Information technology and Gen Z: The role of teachers, the internet, and technology in the education of young people. Technology in Society, 65, 101565.

Talan, T., & Kalinkara, Y. (2022). Students' Opinions about the Educational Use of the Metaverse. International Journal of Technology in Education and Science, 6(2), 333-346. https://doi.org/10.46328/ijtes.385.

Tamilmani, K., Rana, N. P., & Dwivedi, Y. K. (2019). Use of 'habit' is not a habit in understanding individual technology adoption: a review of UTAUT2 based empirical studies. In Smart Working, Living and Organising: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2018, Portsmouth, UK, June 25, 2018, Proceedings (pp. 277-294). Springer International Publishing.

Teng, D., Lai, C., Song, Q., Yang, X., & Liu, L. (2023). Super multi-view near-eye virtual reality with directional backlights from waveguides. Optics express, 31(2), 1721-1736.

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Quarterly, 157-178. https://doi.org/10.2307/41410412.

Yang, F., Ren, L., & Gu, C. (2022). A study of college students' intention to use Metaverse technology for basketball learning based on UTAUT2. Heliyon, 8(9), e10562. https://doi.org/10.1016/J.HELIYON. 2022.E10562.

Yu, H. (2024). Why do people use Metaverse? A uses and gratification theory perspective. Telematics and Informatics, 102110.

Zhai, X. S., Chu, X. Y., Chen, M., Shen, J., & Lou, F. L. (2023). Can Edu-Metaverse reshape virtual teaching community (VTC) to promote educational equity? An exploratory study. IEEE Transactions on Learning Technologies, 16(6), 1130-1140.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Ética, política e informação: uma reflexão conceitual sobre sociedade e comportamento (Ethics, politics, and information: a conceptual reflection on society and behaviour)

Vitor Serejo Ferreira Batista

DOI: https://doi.org/10.47989/ir30iConf47590

Abstract

Introduction. Considering the changes that have occurred in society, from an ethical and political perspective linked to the informational aspects of production, processing, and mediation of information, we propose as our objective a conceptual reflection on society and its behavior.

Method. Our methodology consists of analysing the concepts linked to society, considered as a matter of critical reflection on the way in which social conduct has been presenting changes in relation to its ethical purpose.

Analysis. To this end, we ask ourselves how the use of technological devices, as well as social media, such as the internet and other networks, influence this behavior. Along with this, we point to types of knowledge production that we see emerging in research on Brazilian society, as a debate counter-productive to the legacy of colonialism, b4ased on the notion of popular wisdom.

Results. Our results are based on the hypothesis that, from the knowledge society onwards, our goal is no longer to control information as before, as was the case with the other types of society mentioned, but rather to produce information using these instruments as a source for new information, not necessarily true, fueling a post-truth society that is often not aligned with the values stipulated by the State.

Conclusion. We conclude that the findings of the researchers cited throughout the work are relevant, as a valid proposal for an ethical sense for the treatment of information in Brazil. Thus, the means presented by researchers such as Rufino (2023) and Simas (2024) contribute to a popular and Brazilian notion of an ethical vision of information.

Introdução

Este trabalho se baseia na experiência prestada pelo autor, ainda na primeira metade do ano de 2024, em ocasião de ministrar disciplina de graduação no Departamento de Informação, na Universidade Federal Fluminense (UFF), em Niterói, Rio de Janeiro. No devido momento, a disciplina *Ética e Informação* lhe foi oferecida a partir de uma perspectiva epistemológica pautada na história de elementos éticos ligados à informação, com cunho de elementos presentes nas experiências europeias de reflexão filosófica sobre o comportamento da sociedade.

Ao partir de uma explanação sobre a ética, tendo como base o modelo clássico grego aristotélico, perpassando pelo afastamento da ética da política no renascimento, a aproximação da moral kantiana e os desdobramentos dos modelos éticos ligados à burocracia weberiana e de responsabilidade investigada por Arendt, o questionamento feito na disciplina se fez: quais modelos éticos que hoje nos são apresentados e desenvolvidos no Brasil? Seguimos ainda esse tipo de perspectiva europeia que estudamos?

O intuito destas perguntas, portanto, se baseiam nos resultados obtidos desta reflexão, junto aos alunos da disciplina, sobre a existência de um papel da sociedade brasileira sobre a ética, tendo como ponto de partida a reflexão acerca de alguns fenômenos ligados ao conceito de sociedade, aqui em específico: sociedade do controle, sociedade de informação e sociedade do conhecimento.

Para nossa reflexão neste trabalho, trabalhamos com a análise de conceitos ligados ao campo da informação, entre eles os citados: sociedade de vigilância, sociedade de controle, sociedade da informação e sociedade do conhecimento. Além da proposta deste debate conceitual, também fazemos uso de uma metodologia qualitativa, tendo como método uma revisão bibliográfica sobre estes conceitos, assim como de obras que surgem como propostas de exposição sobre esse saber popular, muito dele provindo das experiências da diáspora.

Compreendemos o saber popular como o conhecimento que emana do povo, de forma a se encontrar nas suas repetições ou nas ações do povo, como rituais de tradição religiosa, tipos de comida e costumes característicos de uma região em específico. Esse tipo de costume, apresentado por Edison Carneiro em A sabedoria popular (2008), é uma fonte de possibilidade para a compreensão do ser no Brasil que antes era conhecido como um ser escravizado.

Sobre os conceitos de sociedade

É curioso pensarmos que uma sociedade se forma por conta da maneira como ela se interage, cria hábitos e costumes, valores e crenças. Todos estes apetrechos citados são movidos e formados por duas vertentes fundamentais: comunicação e informação. Saber comunicar ou como comunicar, fazer uso dos veículos adequados e oportunos é quase uma arte, um jogo de palavras propícias, macete escolhido pelo emissor para com o seu receptor, ou locutor com o interlocutor. No caso da informação, a mensagem é um dilema. Isso porque, segundo Capurro (2007), podemos considerar a informação como um elemento fundamental das sociedades contemporâneas, principalmente a partir da segunda guerra mundial.

A partir desta potência dada à informação, podemos pensar que sua mediação, assim como seu controle, leva a uma formação de sociedade específica, pautada por alguns e difundida para muitos, os quais muitas das vezes não sabem ou não podem mudá-la. Essa afirmação, um tanto romântica e com tons revolucionários não é à toa, mas não procura no leitor um grito de guerra ou revolta. Faz parte da construção social em que se dá uma sociedade que busca nos valores morais e de controle um modelo próprio, porém nada original.

Dito isso, numa perspectiva geral e histórica, podemos compreender o conceito de sociedade como uma estrutura composta pelas instituições e os tipos de relação social presentes em uma comunidade. Proveniente das pessoas que ali se apresentam, tal como os modos de operação social

como interagem entre si, a sociedade tem por si uma espécie de hierarquização social, tendendo a ser dividida. Esse tipo de divisão não só possibilita uma desigualdade entre as pessoas e os grupos que ali estão, como também acena para uma distribuição de valores e regras que a compõem.

Nessa reflexão acerca destes valores sociais que ali estão, se apresenta a ética, como um modo de pensamento que propõe o debate sobre estes costumes limitados por um período de tempo, questões culturais e outros adventos que possam remeter a ideia de moral. O controle das ações e a padronização sobre a ideia de moral, ainda mais aproximada da Ética kantiana (Marcondes, 2005.), em meados do século XVIII, fez com que parte desta sociedade, assim como suas instituições, códigos e dispositivos legais, se tornassem grupos ligados a um poder.

O poder, de maneira geral, possibilita com que uma sociedade seja governada ou se mantenha de forma conservada aos membros que ali se detém no poder. Sendo assim, estabelecer meios de exercer esse tipo de manutenção se fizeram notáveis ao longo dos anos, aqui em particular do século XX. Entre esses meios, destacamos o surgimento das sociedades de controle e vigilância (Deleuze, 1992).

O que antes se fazia perceber em uma sociedade pautada pela disciplina, através de aparelhos ideológicos do Estado (Althusser, 1967), percebemos como um controle destes aparelhos sobre o sujeito que ali se fazia como um ser móvel na sociedade. Sua mobilidade, no entanto, se fazia dentro de uma esfera restrita de atuação, tendo como pauta maior os ambientes característicos de controle social: escola, igreja, família, etc. Até mesmo as rodas de amigos poderiam ser vistos como tal espaço de controle, uma vez que os atos das pessoas se mantinham num mesmo padrão. Cenário um tanto distópico em termos literários, esse tipo de modelo social só se intensificou com o avanço tecnológico, principalmente num cenário eletrônico. Câmeras de vigilância, substitutas dos olhos de uma supervisão humana, podem ser um excelente molde figurativo de uma sociedade de vigilância, bem aos moldes foucaultianos do panóptico desenvolvido por Bentham, ainda no século XIX.

Se pensarmos em modelos mais contemporâneos, esse tipo de fenômeno se dá no controle de dados ou na vigilância de imagens produzidas como desempenho de ostentação nas redes sociais, seja por questões psicológicas ligadas à necessidade de exposição do eu, ou mesmo pela demonstração em registro da experiência. Zuboff (2021), por exemplo, nos demonstra como a figura do sujeito cliente para com a sociedade, ou com quem a controla, expõe as nuances de um capitalismo de vigilância contemporâneo. Os dados, hoje em dia, como a mesma nos lembra em sua obra, são tão valiosos quanto o petróleo em outras épocas.

Da mesma forma, Beiguelman (2021) nos demonstra como a questão dos status como modo de experiência digital, porém um tanto social por razão das redes, produz reflexões na comunicação, nas relações afetivas, assim como nas estruturas de produção documental. Esse tipo de resultados nos aponta para uma mudança estética de vigilância, que não se ocupa somente das faces nas redes, mais se expande em outros ambientes pela cidade, como em ambientes de espetáculos sonoros, tais concertos de música, jogos esportivos, até mesmo festas de rua, como nas festas de final de ano no Rio de Janeiro e no Carnaval.

No sentido mais próximo de reflexões como as de Debord (1973), parecer é um elemento fundamental em uma sociedade que tem a necessidade em se mostrar, mesmo que essa amostra pessoal seja por meio de uma experiência incompleta, talvez mais próxima de uma pobreza, como nos chama atenção Benjamin (1994). Essa ausência de uma essência das atividades, por assim dizer, pode ser compreendida num ambiente em que não compreendemos um porquê em fazer, mas fazemos porque temos de fazer. Em outras palavras, um condicionamento. Em uma sociedade que vivenciou uma experiência colonial e à nega como história, isso se torna curioso por abraçar mais

a experiência do colonizador do que a do colonizado, de se parecer mais com o outro do que se reconhecer como é.

Reflexões acerca da tecnologia e da sabedoria popular

As questões ligadas ao desenvolvimento das tecnologias são, de certo modo, um tanto interessantes no que diz respeito à compreensão do que seria a tecnologia. Hoje em dia, muito assimilada à idéia dos meios digitais e eletrônicos, a tecnologia nada mais é do que uma habilidade aplicada para se desenvolver um modo de fazer algo, um conjunto de processos para se chegar ao objetivo de se saber fazer. Estas formas de aprimoramento do forma em se fazer algo, essa técnica específica, é o que compreendemos por tecnologia (Aulete, 2011.).

O uso de dispositivos eletrônicos, assim como o de programas desenvolvidos para um desenvolvimento de comunicação e integração social em ambiente virtual, se propagou em princípios do século XXI, se tornando um símbolo interessante do futuro desse início de novo século. Através de um recorte de tempo, principalmente no período pandêmico de 2020 em diante, fizemos destes dispositivos um instrumento diário e, praticamente, indispensável na forma de se manter visto.

Sobre a questão do dispositivo, gostaríamos de propor uma breve reflexão sobre o termo, a partir da noção trazida por Agamben (2005), na qual o autor nos apresenta a hipótese que "a palavra "dispositivo"[...] seja um termo técnico decisivo na estratégia do pensamento de Foucault" (Agamben, 2005, p. 9). Esse pensamento de Foucault, citado por Agamben, faz referência a noção foucaultiana de dispositivo de poder, a qual nos desdobramos agora. Segundo o filósofo italiano, são três as possibilidades de atuação, sendo elas:

> 1) É um conjunto heterogêneo, que inclui virtualmente qualquer coisa, linguístico e não-linguístico no mesmo título: discursos, instituições, edifícios, leis, medidas de segurança, proposições filosóficas etc. O dispositivo em si mesmo é a rede que se estabelece entre esses elementos.
> 2) O dispositivo tem sempre uma função estratégica concreta e se inscreve em uma relação de poder;
> 3) É algo geral (uma rede) porque inclui em si a episteme, que para Foucault é aquilo que em certa sociedade permite distinguir o que é aceito como um enunciado científico daquilo que não é científico (Agamben, 2005, p. 9-10).

Nesse ponto, o que fazemos é uma aproximação deste dispositivo mencionado por Agamben, e que remete a Foucault, para o dispositivo eletrønico, tal como as redes sociais, e porque não os aparelhos usados de suporte para esses programas e redes, para debater sobre essa estratégia prevista pelo filósofo italiano. O que podemos perceber é que em meio a um discurso de comunidade e felicidade das redes, o que também temos são milhares de postagens, que remetem a outro número avassalador de informações, com o intuito de promover entre seus usuários uma finalidade de comunicação e visualidade pessoal.

Quando pensamos na questão de "ser visto", como forma de existência, podemos perceber como que a necessidade em se expor, assim como a maneira como nos expomos para ser um dilema para muitos, por conta de sua maneira de lidar com a tecnologia em maior destaque. As redes sociais também se tornaram um chamariz de nos percebermos enquanto pessoas, de forma que muitas pessoas, ao se ausentar das redes, parecem estar se ausentando da vida pública, ou de uma esfera de sociabilidade.

Pensar isso como uma verdade é no mínimo questionável. Mas descartar essa reflexão é no entanto uma falha, no sentido de compreendermos a força pungente por trás das redes sociais. E não só isso, o uso de inteligência artificial, tal como a deepfake e outras formas de produção de conteúdo - muitas das vezes vinculado ao uso da imagem como recurso de informação - vem se tornando algo a ser não só observado, mas estudado.

Este trabalho não vai se debruçar nessas questões tecnológicas, mas nas repercussões que estas vêm causando na sociedade, principalmente na experiência brasileira.
Simas (2024), em sua obra *Umbandas: uma história do Brasil*, nos alerta para esse fenômeno, assim como o resultado de uma experiência brasileira voltado para uma questão cultural que se fortalece ao longo dos anos, da seguinte forma:

> Em tempos cada vez mais marcados por relações mediadas por redes sociais online, capazes de conectar pessoas ou instituições que a princípio partilham interesses, práticas e objetivos comuns, é impressionante a quantidade de conteúdo vinculado às umbandas: giras, consultas e até entrevistas com entidades veiculadas no YouTube, fotos compartilhadas com textos no instagram, postagens no Facebook, fios no Twitter, vídeos curtos no TikTok, pipocam em grande profusão nas redes (Simas, 2024, p.7)

O que podemos perceber na afirmação do professor Simas é a de que as redes sociais são um, se não o melhor veículo comunicativo da nossa era, de forma que difunde temas dos mais variados, e por que não dizer, de importante reflexão a nível nacional, como é o caso da Umbanda no Brasil. Não estamos aqui propondo uma divulgação ou explanação de cunho religioso. Mas o que propomos é uma reflexão a partir de um tipo de pensamento que se forma nesta prática vinculada à uma herança diaspórica africana e que se encontra no Brasil numa encruzilhada de possibilidades.

Ainda sobre isso, o que gostaríamos de pensar aqui não é necessariamente o conteúdo imagético exposto nas redes sociais, mas o fenômeno provocado por elas no quesito difusão desta informação. O papel que antes parecia ser de instituições tais como os arquivos, as bibliotecas, museus e centros de memória, hoje já superou a televisão como veículo de maior alcance, e tem nas redes sociais e na internet seu maior difusor.

Dito isso, em contraponto a produção imagética esplanada, gostaríamos de refletir sobre a necessidade de se formar imagens que não necessariamente são verdade, ou seja, um rompimento para com o que é; um foco para o que parece ou poderia ser. Nesse quesito, o uso das imagens como uma crença na verdade, ou a possibilidade de confirmação de algo em que se pode crer, é o que nos parece mais necessário nesta crítica. Então, por que dependemos tanto das imagens como de um recurso informacional válido?

Assim como o uso da imaginação tecnológica e digital é de grande efeito em nosso tempo, o que propomos aqui é a observação de um recurso tecnológico analógico que, por muitos, parece esquecido: o gesto da tradição das ações, a partir do uso cotidiano de falas, danças, cantos, quaisquer sejam as ações, de forma a transmitir esse recurso como uma sabedoria popular.

A sabedoria popular, enquanto uma forma de produção do conhecimento, capaz de ser coletada, organizada e tratada, é um recurso interessante para os estudos em organização do conhecimento, uma vez que traz consigo uma potencial metodologia popular, a partir de uma série de técnicas e maneiras de fazer, as quais compõem esses saberes como um jeito particular de um pensamento, uma potencial ciência. Na verdade, esse tipo de ação nos parece mais com uma proposta decolonial do conhecimento, com o objetivo de deixar transparecer ações éticas e propostas epistemológicas

de culturas oprimidas, e que no entanto, vem praticando estratégias de troca de informação e afirmação de saberes por meio de publicações físicas editoriais e, como já mencionamos, difusão via redes sociais. Sobre esse processo de descolonização do conhecimento, Garcez e Sales (2021) nos afirmam que:

> A descolonização do conhecimento está presente em diversas abordagens que se colocam contra as perspectivas hegemônicas de produção do conhecimento. Quando nos voltamos para os saberes que se colocam como contracoloniais ou decoloniais, podemos pensar, por exemplo, nas perspectivas oriundas de mulheres, pessoas de origem latina, africana, indígena, ameríndia ou ribeirinha (Garcez; Sales, 2021, p. 7).

Ao pensarmos isso, gostaríamos de expor algumas destas sabedorias populares, as quais nos parecem ser poderosos agentes de mudança na forma como a sociedade se compreende como tal, a ponto de se afirmar mais como um fruto de seu meio do que um reflexo de um processo colonial, a qual provocou na mesma tempos de controle e exploração.

O ser em ginga: Meios de reflexão sobre a informação

Primeiramente, esse conceito, trazido por Rufino (2023) remete a ideia do jogo de corpo, como uma analogia à capoeira, uma relação do corpo que opera em frente dupla, ora em defesa, ora em ataque. Esse movimento provocado na reflexão de Rufino nos faz pensar na construção ética desse conceito, no sentido de provocar um equilíbrio de forças na sociedade.

Para falarmos em sabedorias populares, como as já mencionadas ética da ginga e a ética indígena, precisamos falar em sociedade da informação e do conhecimento. Da mesma forma, precisamos diferenciar um conceito do outro. A sociedade da informação é um fenômeno que surge no século XX, mas que pode ser compreendido graças aos avanços científicos do século XIX, principalmente nas esferas de levantamento de informações sobre o mundo, Nesse sentido, Paul Otlet pode ser considerado um agente valioso nesse aspecto, uma vez que foi quem procurou desempenhar um papel de mapear o mundo (Juvêncio, 2021). Tal estruturação visada por Otlet, porém, compreendia um mundo que se limitava aos interesses europeus.

Percebendo isso, classificar uma sociedade como "da informação", é no mínimo arbitrário, uma vez que você delimita os atores que ali se inserem nesta sociedade, neste grupo global. Em outras palavras, mesmo que outros países sejam trazidos para esse debate, eles mais parecem marionetes periféricos dos outros, comandantes da festa. Tendo isso em mente, Mattelart (2006), nos alerta sobre o seguinte cenário:

> Na produção do imaginário em torno desse futuro balizado, a tecnologia e o discurso de acompanhamento que a serve ocupam uma posição decisiva e singular. O mesmo vale para as apologias da "sociedade da informação", promovida a paradigma dominante da mudança e caução de um mundo "mais solidário, transparente, livre, igualitário". Ora, a sociedade prometida atribuída à nova matéria-prima imaterial só evidencia seu sentido em uma configuração geopolítica (Mattelart, 2006, p. 7).

Num sentido mais interessante, porém também questionável, o surgimento da sociedade de conhecimento possibilita alternativas interessantes sobre essa mudança de cenário globalizado, promovida pela sociedade da informação. O conhecimento, segundo Lyotard (2021), se apresenta como uma das principais forças de produção nos últimos tempos, seja em questões científicas, de trabalho ou questões pessoais. Numa possibilidade de compreensão mesmo pós-moderna, o enriquecer do conhecimento, assim como seu controle, é um mecanismo interessante, uma vez que abre brechas para outros conhecimentos que até agora pareciam não avistados.

Para Mattelart (2005), esse conhecimento nos é apresentado, principalmente, por meio da documentação produzida. A ideia de documentação é objeto de reflexão mesmo em tempos de Otlet, seja pelo Tratado da Documentação (2018), seja por meio das contribuições de Suzanne Briet (2016) ainda nos anos de 1950. A forma como o documento se torna um elemento de validade histórica, seja em âmbito da memória ou recurso jurídico, faz com que possamos pensar nele, o documento, como um objeto válido na construção de cum conhecimento. É aí que nos perguntamos: mas quem produz esse documento? O discurso produzido pelo documento é ponto fundamental para compreendê-lo. Exatamente nesse quesito é que gostaríamos de pensar na ideia de *sabedorias de fresta*, numa forma de entrelinhas do que se pode considerar oficial por uma linha de cultura, e inserir outros tipos de conhecimento no debate. Rufino (2019) nos aponta as sabedorias de fresta como "Mandinga, incorporação, ginga, negaça, transe, rolê, efó, amarração, feitiço, terreiro, esquiva, drible [...] marcas que tecem esse inventário assente nos limites do corpo" (Rufino, 2019, p. 68).

Apesar de mencionarmos a questão da sociedade nesse trabalho, no intuito de adentrarmos um território mais específico do pensamento aqui proposto, utilizamos o conceito de campo, de Bourdieu, mais específico o conceito de campo social que o mesmo apresenta. Na composição de campo para Bourdieu (2008), um campo pode se apresentar de maneiras diferentes, como metáforas que fazem analogia a questões ligadas, por exemplo, a um jogo. Aqui não compreendemos o jogo como uma partida ou uma disputa esportiva, na qual temos alguém contra alguém. Mas sim uma sugestão de agentes que ocupam espaços de forma limitada. Esses agentes seriam, por exemplo, pessoas e instituições.

O que gostaríamos de pensar aqui é num sentido de jogo de equilíbrio, um jogo de capoeira, na qual o campo se converte em uma espécie de roda e os agentes que estão presentes jogam, no sentido de não provocar o desequilíbrio do jogo, ou seja, seu término. Essa ação é possível graças ao movimento que se dá na roda. O movimento do corpo, o canto, a dança, ou seja, as técnicas que envolvem o jogo. Esse movimento, potência para um equilíbrio, atua como um *meio justo*, quase como um pensamento aristotélico sobre a ética. Segundo Rufino (2023), podemos entender esse processo de uma forma educacional, como fundamento corporal e prática mandingueira. Diz o seguinte:

> Os discípulos de Aristóteles ficaram conhecidos como peripatéticos. A palavra, que no grego compreende a noção de itinerante ou ambulante, fazia menção à maneira como o filósofo ensinava caminhando ao ar livre. A aposta da educação aristotélica tinha alguns princípios, um deles destaca a defesa de que a principal função da educação seria conduzir as pessoas à felicidade. A palavra condução como traço de uma aprendizagem itinerante e o processo educativo como parte de um contínuo movimento que demanda corpo enredam o que ficou famoso na relação entre o filósofo e seus discípulos, as múltiplas formas de educar presentes em outras culturas (Rufino, 2023, p. 78).

Perceber a educação aristotélica aproximada ao tipo de reflexão trazido por Rufino, nos faz debater sobre a felicidade como objetivo fim da reflexão ética, trazida pelo grego em sua obra Ética à Nicômaco (1991). Ao trazer para a ideia de virtude como a ação que possibilita uma visão ética sobre as coisas, Aristóteles propõe esse meio justo, um equilíbrio sobre as decisões. Esse mesmo tipo de ação pode ser reconhecida nas rodas de capoeira como a possibilidade em deixar visível o movimento da roda, que tem justificativa porque gira, e não porque para. Assim, proporciona o movimento, o canto, a dança, ou seja, os elementos da tradição da capoeira. É um pensamento, um conhecimento, um jogo.

Nesse mesmo pensamento, Rosa (2013) nos apresenta o conceito de Pedagoginga, como um tipo de compreensão educativa através da ginga, da estética da ginga. Diz o pensador mineiro o seguinte:

> a miragem da Pedagoginga é firmar no fortalecimento de um movimento social educativo que conjugue o que é simbólico e o que é pra encher a barriga, o que é estético e político em uma proposta de formação e de autonomia, que se encoraje a pensar vigas e detalhes de nossas memórias, tradições, desejos [...] (Rosa, 2013, p. 15).

Esse campo, tal como uma roda, tem determinado espaço de atuação e que seus agentes interagem e compõem essa estrutura. A roda se torna um espaço de educação, que gira na busca pelo conhecimento do equilíbrio, da continuidade da ação. Mas apesar disso, Sodré (1988) nos alerta sobre esse conhecimento que se estabelece:

> o mestre capoeirista negro não *ensina* a seu discípulo — pelo menos de maneira como a pedagogia ocidental entende o verbo *ensinar*, ou seja, o mestre não verbaliza nem conceitua o seu saber para doá-lo metodicamente ao aluno. Também não interroga, nem decifra. Ele *inicia*: cria as condições de aprendizagem, (forman-do a roda de capoeira) e assiste a elas. É um proces-so sem qualquer intelectualização, em que se busca um reflexo corporal comandado, não pelo cérebro, mas por algo *indeterminado* resultante dessa inicia-ção do corpo (Sodré, 1988, p. 212).

Esse algo indeterminado, mencionado por Sodré, vem pela experiência na roda, na mandinga mencionada por Rufino (2023), mas principalmente na construção da ação no vazio, ou seja, na ausência deixada pelo colonialismo, num sentido de apagamento da memória, da tentativa de esquecimento das coisas (Pollak, 1989). Apenas pelo ato de repetição, de afirmação do que se experimenta é que poderíamos ter uma noção de realidade, numa espécie de afirmação racional.
Simas (2021) expõe o pensamento de que o Brasil descobriu no vazio a possibilidade da vida existir. Ou seja, no vão das coisas, o movimento se estabelece como um preenchimento do vazio. Isso pode ser identificado, por exemplo, no movimento do capoeira, no drible do futebol, na ocupação do espaço etc. Podemos compreender esse preenchimento dos espaços vazios também como uma noção de sentido, como uma compreensão de ser sobre aquele objeto. Essa compreensão, trazida através de uma prática, cabe aos saberes populares.

O que percebemos é um avanço na documentação, no registro desses saberes, de forma a construir um conhecimento que, segundo Nascimento (2019), tem que ser nomeado, tem de ser falado, tem de ser transmitido Assim como é necessária essa iniciativa e continuidade epistêmica, tais filosofias, que também são formas de compreender políticas e éticas sobre uma sociedade, não dependente de uma outra filosofia, principalmente a do colonizador. Obras como as usadas aqui, seja por Luiz Rufino, Luiz Antônio Simas, Abdias Nascimento, Muniz Sodré, entre outros aqui citados, são construções de uma sabedoria popular brasileira, que como afirma Ailton Krenak e Davi Kopenawa (2021), não fazem do Brasil ainda uma nação, mas são parte de uma composição complexa a qual o Brasil se desdobra e se compõe.

Resultados

Concluímos que a Ciência da Informação, nesse sentido, tem papel interessante na pesquisa sobre esse tipo de prática e de reflexão científica, no que tange uma espécie de ação analítica sobre como o controle sobre da informação talvez não seja a única resposta para compreendermos um futuro da sociedade. Isso porque a noção de documentar um conhecimento, assim como preservá-lo, faz com que ele se faça presente no debate sobre uma sociedade. As hipóteses aqui trazidas são vias destas tentativas.

A Ciência da Informação, até pelo seu caráter interdisciplinar (Borko, 1968) faz com que possamos vê-la não como uma ciência dura, como podemos perceber em outras disciplinas, tal como a própria filosofia e a ética, mas sim como um campo de inter relações, as quais possibilitam a mesma trazer para perto um objeto de pesquisa e desdobrá-lo, como um mosaico de possibilidades de análise científica. Esse trabalho se fez possível nessa perspectiva.

O que podemos constatar na questão da questão da ética como uma base de reflexão e debate filosófico, foi uma ausência de aplicação maior sobre as próprias demandas da sociedade brasileira, frente a sua formação histórica e cultural, devido principalmente a sua experiência de colônia. Visto isso, esse trabalho propõe não somente uma reflexão ética e política dos fluxos de informação e do comportamento nessa sociedade, tendo como ponto de inflexão os conceitos ligados à sociedade.

Junto a isso, se faz presente também uma proposta de analisar uma possível filosofia popular brasileira, que surge destas experiências, e que vem sendo objeto de pesquisa por parte de educadores, historiadores e filósofos brasileiros, mais precisamente da região metropolitana do Rio de Janeiro. Esse tipo de pensamento flerta bastante com a construção da figura boêmia carioca, mais precisamente do malandro, por razão do jogo de corpo que apresenta na sua forma de se expor.

O uso desse tipo de figura, o capoeira, como uma analogia metafórica sobre a maneira como o Brasil poderia lidar com a reflexão ética e política sobre a construção de sua sociedade, não é de fato uma novidade. Não trazemos esse exemplo como um ato de originalidade temporal. Mas percebemos uma crescente desse tipo de reflexão, ou mesmo um retorno a essa estratégia de pensamento, frente a sua associação com a sociedade de conhecimento. Essa *fresta* promovida por esse tipo de momento histórico possibilita tal iniciativa dos pensadores brasileiros.

Caetano Veloso na música *Jeito de Corpo* (1981), ao iniciar a canção, menciona que "*Eu 'to fazendo saber/ Vou saber fazer tudo de que eu sou a fins*". Essa mensagem do compositor baiano é uma síntese do que procuramos revelar nesse trabalho, que buscou apresentar as propostas de sabedoria popular que, através de trabalhos científicos, publicações de livros, palestras etc, vem sendo produzida no território brasileiro. Utilizamos a capoeira como um recurso imagético, característico do Brasil e que vem de herança da diáspora africana.

O uso da corporeidade não é à toa neste trabalho, uma vez que o corpo é tratado, ao longo da história, como um campo pertencente ao movimento das emoções, dos instintos, ou seja, um objeto de controle das ações e das narrativas populares. Domar um corpo, masculino ou feminino, é promover a este limites de espaço. Mestre Pastinha, uma das principais vertentes da Capoeira Angola praticada no Estado da Bahia, dizia em seu Caderno Alba, um manuscrito de filosofia popular, o seguinte: "*Amigos o corpo é um grande sistema de razão, por detrás de nossos pensamentos acha-se um senhor poderoso, um sábio desconhecido.*" (Decanio Filho, 1997).

Podemos dizer que o que antes era desconhecido, segundo Mestre Pastinha, vem ganhando corpo enquanto pensamento, ação, conhecimento disseminado, e que é capaz de mudar sociedades. Da mesma forma como se aproveitou das frestas deixadas pela sociedade do conhecimento, a

sabedoria popular é um recurso, senão uma interessante via, de se compreender melhor um Brasil que se manteve nas frestas, por tempo demais.

Sobre o autor

Vitor Serejo é doutorando do Programa de Pós-Graduação em Ciência da Informação (PPGCI/UFF), onde pesquisa os arquivos pessoais de escritores. Tem experiência na disciplina Ética e Informação, que leccionou anteriormente. Ele se interessa por outros temas, como estudos de arquivo, produtores de arquivos, literatura, heteronomia e autoficção. Pode ser contactado através do endereço vitorserejo@id.uff.br

Referências

Agamben, Giorgio. (2005). O que é um dispositivo? Outra travessia, n. 5, p. 9-16

Althusser, Louis. (1985). Aparelhos ideológicos de Estado. Rio de Janeiro: Graal, v. 2

Aristóteles. Ética a Nicômaco. (1991). In:Os Pensadores. Tradução de Leonel Vallandro e Gerd Bornheim da versão inglesa de W.D. Ross. São Paulo:Abril Cultural.

Aulete, Caldas. (2011). Novíssimo Aulete Dicionário contemporâneo da língua portuguesa; organizador Paulo Geiger. Rio de Janeiro: Lexikon.

Beiguelman, Giselle. (2021). Políticas da imagem: vigilância e resistência na dadosfera. Ubu Editora.

Benjamin, Walter. (1994). Experiência e pobreza. In: Magia e técnica, arte e política. São Paulo: Brasiliense.

BORKO, Harold. (1968). Ciência da Informação: o que é isto. American Documentation, v. 19, n. 1, p. 3-5.

Bordieu, Pierre. (2018). Pierre Bourdieu: conceitos fundamentais/ editado por Michael Grenfell; tradução de Fábio Ribeiro. Petrópolis, RJ: Vozes.

BRIET, Suzanne. (2016). O que é a documentação?. Brasília, DF: Briquet de Lemos.

Caetano Veloso. (1981) Jeito de corpo: Universal Music Group..

Disponível em: https://www.youtube.com/watch?v=3siR1bK7Gws Acesso em: 26 jun. 2024

Carneiro, Edison. (2008). A sabedoria popular. São Paulo, Martins Fontes, 180 pp.

Capurro, Rafael; Hjørland, Birger. (2007) O conceito de informação. Perspectivas em ciência da informação, v. 12, p. 148-207.

Debord, Guy. (1997). A sociedade dos espetáculos. Trad. Estela dos Santos Abreu. Rio de Janeiro: Contraponto.

Decanio Filho, Angelo. (1997) Manuscritos e desenhos de Mestre Pastinha: A herança de Pastinha. Salvador: o autor. 2ª ed.

Deleuze, Gilles. (1992). Post-Scriptum: sobre as sociedades de controle. In: Conversações: 1972-1990. São Paulo: Editora 34. p. 219-226.

Garcez, DC; Sales, R. (2021). Descolonizando a organização do conhecimento. Tendências da Pesquisa Brasileira em Ciência da Informação , v. 14,.

Juvêncio, Carlos Henrique. (2021). Arquitetura das ideias: Paul Otlet, o objeto, o livro e o documento. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 26, p. 01-17.

Lyotard, Jean-François. (2021). A condição pós-moderna; tradução Ricardo Corrêa Barbosa. 20ª ed. Rio de Janeiro: José Olympio.

Marcondes, Danilo. (2005). Iniciação à história da filosofia: dos pré-socráticos a Wittgenstein. 9ª ed. Rio de Janeiro: Jorge Zahar.

Mattelart, Armand. (2006). História da sociedade da informação. 2ª ed. São Paulo: Edições Loyola.

Mattelart, Armand. (2005). Sociedade do conhecimento e controle da informação e da comunicação. Encontro Latino de Economia Política da Informação, Comunicação e Cultura , v. 1-22.

Nascimento, Abdias. (2020). O quilombismo: Documentos de uma militância Pan-Africanista. São Paulo: Editora Perspectiva S.A.

Otlet, Paul. (2018). Tratado de documentação: o livro sobre o livro teoria e prática. Brasília: Briquet de Lemos / Livros.

Pollak, Michael. (1989). Memória, esquecimento, silêncio. Revista estudos históricos, v. 2, n. 3, p. 3-15.

Radio Companhia. (2021).. Ailton Krenak e Davi Kopenawa - O encontro. 7 nov. de 2021. Podcast Disponível em: https://www.companhiadasletras.com.br/blogDaCompanhia/Post/6268/radio-companhia-166-ailton-krenak-e-davi-kopenawa-o-encontro Acesso em: 26 jun. 2024.

Rosa, Allan da. (2013). Pedagoginga, autonomia e mocambagem. Rio de Janeiro: Aeroplano.

Rufino, Luiz. (2019). O que pode Elegbara? Filosofias do corpo e sabedorias de fresta. Voluntas: Revista Internacional de Filosofia, v. 10, p. 65-82.

Rufino, Luiz. (2023). Ponta-Cabeça: educação, jogo de corpo e outras mandingas. Rio de Janeiro: Mórula Editorial.

Simas, Luiz Antonio. (2021). Maracanã: quando a cidade era terreiro. Rio de Janeiro: Mórula Editorial.

Simas, Luiz Antônio. (2024). Umbandas: uma história do Brasil. 9ª ed. Rio de Janeiro: Civilização Brasileira.

Sodré, Muniz. (1988). A verdade seduzida, 2.ed. Rio de Janeiro: Francisco Alves Editora.

Zuboff, Shoshana. (2021). A era do capitalismo de vigilância. Editora Intrínseca, 2021.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Evaluating techniques of artificial intelligence for social robots

Jessica K. Barfield
DOI: https://doi.org/10.47989/ir30iConf47584

Abstract

Introduction. Based on techniques of artificial intelligence (AI) robots equipped with algorithms are becoming more social and intelligent as they enter society. As a relatively unexplored topic of research the current study evaluated whether the perception of robot intelligence was influenced by different techniques of AI.

Method. In an online study participants viewed two versions of a humanoid robot which varied by their surface colour and stated AI abilities. For each image participants rated the perceived intelligence of the robot.

Results. Using an online survey, the results found no statistically significant effect for robot surface colour on judgments of robot intelligence but that robot voice enablement and the ability to detect a user's face and emotions added significantly to the perception of robot intelligence. In addition, amongst the AI techniques evaluated, text used for human-robot communication was the least effective method for conveying the perception of intelligence for a humanoid robot.

Conclusion. As tentative conclusions, the perception of robot intelligence can be based on the specific AI technique used to design the robot, and it appears that the more the human-like AI ability of the robot, the more likely that users will view the robot as intelligent.

Introduction

We live in an age of algorithms in which the technologies that we interact with are gaining in intelligence through neural nets, machine learning techniques, and the use of different statistical procedures. In this paper I take a broad view of intelligence and propose that it reflects a machine or software's ability to perform tasks that are typically associated with human intelligence. One type of intelligent technology that is entering society is social robots which may operate using the latest techniques of artificial intelligence (AI). For example, using computer vision robots can navigate the environment and avoid obstacles (Russell & Norvig, 2020), and using machine learning algorithms robots can detect a person's face, and to some extent, their emotional state (Liu et al., 2017). Robots also have the ability to communicate with users through text or spoken language (Edwards et al., 2019). Surprisingly, with the increased use of algorithms creating smart technologies, there has been relatively little research to determine the effectiveness of AI-driven techniques from a usability perspective. To address this gap in the literature this paper explored how people evaluated different techniques of AI when used to guide the behaviour of a social robot.

As background, past studies have shown that different AI abilities controlling a robot's performance can influence the user's perception of the robot and its problem-solving ability (Flores-Fuentes et al., 2014; McKee, 2003). For example, Dou et al. (2021) found that the use of a robot equipped with natural language processing was effective for shopping, education, and as a home companion. Further, Makibuchi et al. (2010) showed that an algorithmic-based problem-solving approach combined with a humanoid robot was effective for solving problems encountered in a real-world setting. In a different application, Chen et al. (2021) investigated the effectiveness of AI for customer experiences when chatbots were used. Using a quantitative approach, they found that the usability of the chatbot had a positive influence on extrinsic values of customer experience, whereas the responsiveness of the chatbot had a positive impact on intrinsic values of customer experience.

Evaluating another AI technique, Roundtree and Moallem (2021) investigated the use of facial recognition ability for young adult and adolescent populations. In a review of the literature they concluded that facial recognition raised important questions about the clinical and ethical applications of the technology. They also noted that prior studies had not adequately addressed the complexities of facial change over time and for different ethnicities thus influencing the accuracy of facial recognition software (Roundtree & Moallem 2021). They concluded that while facial recognition is a promising technology, usability studies were mostly lacking for facial recognition applications. For a different application, Riaz et al. (2022) looked at the use of facial recognition as a biometric platform and particularly in the context of employment in a real-world setting. Specifically, they evaluated the usability and vulnerability of biometric technologies implemented in the UAE public transportation system to boost security and public service delivery. They noted that usage of facial recognition technology had raised numerous concerns about biometric data security and privacy. Based on participants responses collected using a survey instrument they concluded that public transport users had a poor impression of facial recognition technology compared to iris recognition and fingerprints authentication.

Considering the usability of robots, in addition to the use of AI to control the behavior of robots, the physical design of the robot has been shown to influence user evaluations of the robot. For example, using the shooter bias paradigm, Bartneck et al. (2018) found that people would shoot faster at a robot that had a darker surface colour than a lighter colour robot. Additionally, Sparrow (2020) commented that people racialized a robot based on its design, and particularly surface colour. Barfield (2021) also found that a robot's surface colour influenced the task that robots would be selected to perform with lighter-coloured robots selected for tasks which required problem solving ability and darker-coloured robots selected for tasks which were more labor intensive.

From the above brief discussion, missing in the literature are studies which looked at features of a robot's design in the context of evaluating AI techniques guiding the robot's performance. This led to the following question addressed in the current study- would the sur-face colour of a robot, thought by Sparrow (2020) to racialize a robot, influence its evaluation when equipped with different techniques of AI? More specifically, based on the above discussion of the literature, in this study the primary objective was to evaluate an individual's perception of robot intelligence as a function of different AI techniques controlling the robot. To reiterate, the main interest was to determine if robot surface colour influenced the evaluation of robot intelligence when combined with different AI abilities. The research is guided by the following research questions.

> RQ1: Will robot surface colour (black or white) influence the perception of robot intelligence?
>
> RQ2: Will different techniques of AI result in the same or different ratings of robot intelligence?

Methodology

Experiment design

After obtaining IRB approval, 27 participants (21 male, 6 female) whose mean age was 32.25 years were recruited from online social media sites to participate in the study. The study used two levels of robot surface colour (black, white), and four descriptions of robot intelligence based on the use of different AI techniques (see Table 1). As protocol, participants viewed all AI conditions and robot colours presented in a random order for each participant. After giving consent to participate in the study participants viewed the robot images and then completed an online questionnaire rating the intelligence of the robot just viewed. Further, after viewing all robot conditions participants answered questions rating robot intelligence considering the different AI techniques presented with the robots.

Procedure, dependent variable, and robot narrative

The dependent variable consisted of answers to an online survey which used 1-7 Likert items to evaluate perceived robot intelligence. When the robot used voice to communicate, the participant heard the robot speak a narrative indicating the robot (Figure 1) and user would be working together to solve problems. Each narrative supported the experiment protocol by informing the participants of the AI techniques associated with a particular robot. For the non-voice condition, the narrative was placed on the screen in text and read by the participant. In either case, the robot appeared on the screen while the participant read the narrative or heard the narrative spoken by the robot. The text to speech software used to produce the robot voice was TTSApp for Visual Basic. The robot voice was gendered male and the language spoken was English. The audio format supported was 16kHz, with 16-bit sampling frequency. old heading, normal text.

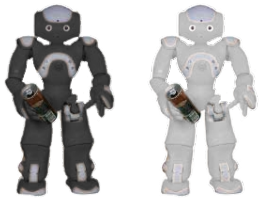


Figure 1. The two robots used in the study

The narrative spoken by the robot allowed the robot to inform the user of the AI techniques it was equipped with (basically low or high levels of AI crossed with the two robot colours). For example, the narratives for the AI condition with the most AI techniques indicated that the robot had motion and collision detection abilities, the capacity for facial and emotion detection, was mobile, and had

natural language processing ability. When the narrative excluded these capabilities and the robot communicated with text only (Barfield, 2021) the narrative described the least amount of AI techniques. Table 1 shows the four narratives used to prime the AI condition. Thus, for each condition a white or black robot appeared on the participant's screen and the narrative indicting different AI abilities was either read by the participant or spoken by the robot.

AI Robot Enabled Condition with Voice	Hi, I'm a robot and I will be working with you to solve a number of problems that we have been assigned to work on together. I hope you find my help useful. I have the following abilities: I can recognize your face and emotions with facial recognition and emotion recognition software, I am mobile and can avoid obstacles, I can understand your speech and talk back to you with natural language ability.
AI Robot Less Enabled Condition with Voice	Hi, I'm a robot and I will be working with you to solve a number of problems that we have been assigned to work on together. I hope you find my help useful. I have the following abilities: I can understand your speech and talk back to you with natural language ability.
AI Robot Enabled Condition without Voice	Hi, I'm a robot and I will be working with you to solve a number of problems that we have been assigned to work on together. I hope you find my help useful. I have the following abilities: I can detect your face and emotions with facial recognition and emotion recognition software, I am mobile and can avoid obstacles. I can understand what you text to me and I can text you back.
AI Robot Less Enabled Condition without Voice	Hi, I'm a robot and I will be working with you to solve a number of problems that we have been assigned to work on together. I hope you find my help useful. I can understand what you text to me and I can text you back.

Table 1. The AI narratives for each of the two robot colours

Results and discussion

The following is a preliminary analysis of the data as further analysis are being performed. Participants were asked to rate robot intelligence using 1-7 Likert items, from this an ANOVA was performed on robot surface colour crossed with the different types of AI enablement presented in the four narratives.

For judgment of robot intelligence, the ANOVA procedure indicated that the main effect for robot surface colour was not statistically significant ($p > .05$). Thus, there was no significant difference among participants in the perception of robot intelligence as a function of whether the participant viewed a black or white-coloured robot. However, the main effect for the different AI enablement narratives was highly significant ($p < .001$). The Tukey HSD multiple comparison test indicated that voice enablement resulted in the perception of a more intelligent robot than a robot that communicated with text ($p < .01$). Further, facial recognition ability led to the perception of more perceived intelligence than text communication ($p < .01$), and similarly, the ability to detect the user's emotions was evaluated as more useful for judging robot intelligence than the ability to communicate with text ($p < .01$). Interestingly, the two-way interaction between AI enablement and robot surface colour was not statistically significant ($p > .05$); indicating that the black or white colour of the robot did not affect the evaluation of the robot's intelligence when the different AI techniques were considered.

Another question asked how intelligent robots were thought to be (with black and white robots combined in one group) when considering the AI techniques of voice communication, facial, and emotion recognition. A one-way ANOVA with voice, facial recognition, and emotion recognition abilities was not statistically significant ($p > .05$). More broadly, Figure 2 shows a graphical depiction of the participant's evaluation of robot intelligence as a function of voice or text communication, and facial and emotion recognition ability. Supporting the results of the multiple comparison test

presented earlier, the figure shows that text communication was rated lowest in the evaluation of robot intelligence.

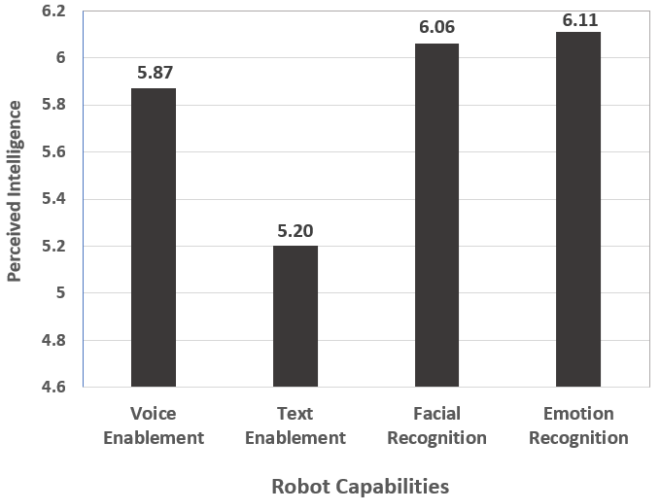


Figure 2. Robot intelligence based on AI techniques

Overall, the results indicated that emotion detection ability, facial recognition, and natural language ability increased the ratings of robot intelligence comparted to text communication. However, it is interesting to note that these AI techniques were not judged to be significantly different from each other when participants evaluated robot intelligence. But consistently, participants evaluated a robot that communicated with text as being less intelligent. It was also shown that colorising a robot as black or white did not affect the evaluation of robot intelligence. In fact, across the different types of AI enablement, the evaluation of robot intelligence was remarkedly similar for white and black-coloured robots (Figure 3). old heading, normal text.

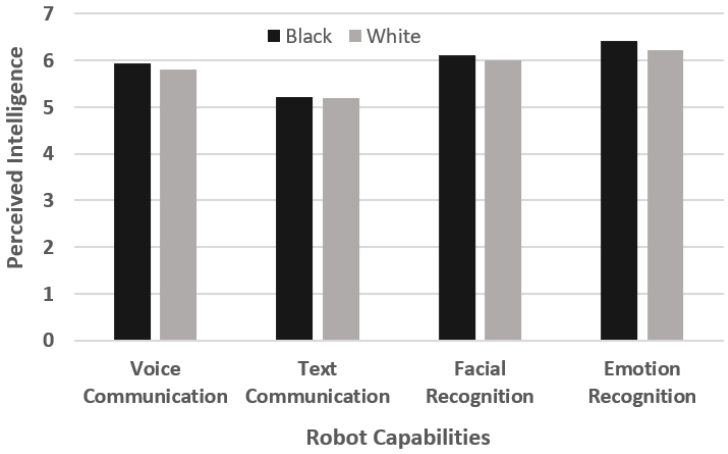


Figure 3. Perceived robot intelligence as a function of robot colour

As a discussion point, I should note that these results are based on the viewing of a static robot image and not based on interacting with a robot to perform a task; for that reason, I postulate that the type of task performed could influence the participants rating of robot intelligence and will be the topic of a future study (Ciocirlan et al 2019). As another point to emphasize, the current study

should be considered exploratory in that this study reflects a beginning effort by the author to investigate a relatively unexplored topic: whether different techniques of AI influence the user's evaluation of a social entity. In the current study, the use of voice was a strong cue to increase the perception of robot intelligence and with large language models beginning to be used with social robots, going forward, it is likely that robots will be perceived as an even more intelligent entity during social interactions. Thus, voice communication seems to be a factor in usability for social robots which leads to another point- the perception of intelligence for an artificial entity could be an important factor to consider when evaluating the usability of an AI-equipped entity. From this observation perhaps a robot intelligence scale should be developed and used in usability studies. It was also an interesting finding that robot surface colour had no effect on the perception of robot intelligence. If robots are racialized as indicated by Sparrow (2020a) it is possible that racial stereotypes may not be a factor in the evaluation of robot intelligence. This conclusion warrants further exploration. Finally, given the viewing of a static robot image, facial and emotion recognition, robot mobility, and natural language ability were not different in user evaluations of robot intelligence. Whether the results presented here are replicated in future studies using different robots and tasks, and operating with more interactivity between user and robot represents a future direction of my research. old heading, normal text.

About the author

Jessica K. Barfield is Assistant Professor in School of Information Science and the Human-Robot Interaction Lead Faculty of the Intelligent Robotic Arms (IRA) lab in the College of Engineering, University of Kentucky, Lexington, Kentucky, USA. She can be contacted at jessicabarfield@uky.edu.

References

Barfield, J. K. (2021). Discrimination and stereotypical responses to robots as a function of robot colorization. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct)*. Utrecht, the Netherlands.

Bartneck, C., Yogeeswaran, K., Sparrow, R., Wang, S., & Eyssel, F. (2018). Robots and racism. In *ACM/IEEE International Conference on Human-Robot Interaction* (pp. 196–204).

Chen, J.-S., Le, T.-T., & Florence, D. (2021). Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing. *International Journal of Retail & Distribution Management*, 49(11), 1512-1531.

Ciocirlan, S., Agrigoroaie, R., & Tapus, A. (2019). Human-robot team: Effects of communication in analyzing trust. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1-7).

Dou, X., Wu, C. F., Linz, K. C., Gan, S. Z., & Tseng, T. M. (2021). Effects of different types of social robot voices on affective evaluations in different application fields. *International Journal of Social Robotics*, 13(4), 615-628.

Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: Social identity theory in human-robot interactions. *Computers in Human Behavior*, 90, 357-362.

Flores-Fuentes, W., Rodriguez-Quinonez, J. C., Hernandez-Balbuena, D., et al. (2014). Machine vision supported by artificial intelligence, applied to rotatory mirror scanners. In *IEEE 23rd International Symposium on Industrial Electronics (ISIE)* (pp. 1949-1954).

Liu, Z., et al. (2017). A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4), 668-676.

Makibuchi, N., Shen, F. R., & Hasegawa, O. (2010). Online knowledge acquisition and general problem solving in a real world by humanoid robots. In K. Diamantaras, W. Duch, & L. S. Iliadis (Eds.), *ICANN, Part III, LNCS 6354* (pp. 551–556). Springer-Verlag Berlin Heidelberg.

McKee, G. T. (2003). An online robot system for projects in robot intelligence. *International Journal of Engineering Education*, 19(3), 356-362.

Riaz, S., Mushtaq, A., & Ibrar, H. (2022). Analyzing and comparing public perception of facial recognition, iris verification and fingerprints based authentication systems. In *International Conference on Control Decision and Information Technologies* (pp. 641-646).

Roundtree, A. K., & Moallem, A. (2021). Testing facial recognition software for young adults and adolescents: An integrative review. In *HCI for Cybersecurity, Privacy and Trust (HCI-CPT 2021)*, 12788, 50-65.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach*. Pearson.

Sparrow, R. (2020). Do robots have race? Race, social construction, and HRI. *IEEE Robotics and Automation Magazine*, 27(3), 144-150.

Sparrow, R. (2020a). Robotics has a race problem. *Science Technology and Human Values*, 45(3), 538-560.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Taking disagreements into consideration: human annotation variability in privacy policy analysis

Tian Wang, Yuanye Ma, Catherine Blake, Masooda Bashir, and Ryan Wang
DOI: https://doi.org/10.47989/ir30iConf47581

Abstract

Introduction. Privacy policies inform users about data practices but are often complex and difficult to interpret. Human annotation plays a key role in understanding privacy policies, yet annotation disagreements highlight the complexity of these texts. Traditional machine learning models prioritize consensus, overlooking annotation variability and its impact on accuracy.

Method. This study examines how annotation disagreements affect machine learning performance using the OPP-115 corpus. It compares majority vote and union methods with alternative strategies to assess their impact on policy classification.

Analysis. The study evaluates whether increasing annotator consensus improves model effectiveness and if disagreement-aware approaches yield more reliable results.

Results. Higher agreement levels improve model performance across most categories. Complete agreement yields the best F1-scores, especially for First Party Collection/Use and Third-Party Sharing/Collection. Annotation disagreements significantly impact classification outcomes, underscoring the need for understanding annotation disagreements.

Conclusion. Ignoring annotation disagreements can misrepresent model accuracy. This study proposes new evaluation strategies that account for annotation variability, offering a more realistic approach to privacy policy analysis. Future work should explore the causes of annotation disagreements to improve machine learning transparency and reliability.

Introduction

Privacy policies are crucial for digital privacy, explaining how businesses handle personal data to help users make informed decisions. However, the FTC and other studies find most are '*incomprehensible*', with users rarely reading or understanding them (Braun, 2024; Hallgren, 2012; Tang et al., 2021). Technical language and inconsistent terminology further complicate comprehension (Azhagusundari & Thanamani, 2013), and ambiguities lead to varying interpretations of terms (Pedregosa et al., 2011). Users' understanding also depends on their education and cultural background (Hossin & Sulaiman, 2015; Krippendorff, 2018).

Researchers have used natural language processing (NLP) to improve the readability of privacy policies (Li, et al;, 2022; Wilson et al., 2016) Efforts include summarising lengthy policies (LingPipe Alias-i., 2008), using topic models to extract data practices (Harkous et al., 2018), classifying content (Grosman et al., 2020), identifying specific information (Gray, 2011; Mysore Sathyendra et al., 2017) and checking compliance automatically (Liu, et al., 2016). Using machine learning (ML) and natural language processing (NLP) to improve privacy policy readability offers benefits, such as processing large data quickly and outperforming human annotators in speed. However, human annotation, often used to create gold-standard datasets, is time-consuming and relies on inter-rater reliability, leaving disagreements unexplored. ML models, which optimise probability-based classification, prioritise data quantity over annotation variety (Plank, 2022). This overlooks data inconsistencies and diverse interpretations. A study found that even privacy experts rarely agree on policy interpretations, suggesting automated tools may struggle to accurately interpret policies, just like users (Braun, 2024).

Human annotation and interpretation of policy and legal texts are often inconsistent and prone to disagreement. For instance, annotators assigned three different labels—Other, User Choice/Control, and First Party Collection/Use—to the same text, while another sentence received two distinct labels: first and third party, and User Choice/Control. Similarly, texts (4)-(7) were inconsistently labelled as First Party Collection/Use by some annotators but not by others. This variation stems from the deliberate ambiguity and complexity of such texts, reflecting their inherently controversial and evolving nature.

Texts to be annotated/ labelled:

1. If you do not wish to share your PIN, you always have the option to not provide the information or use the MediaNews Websites that require it.
2. By use of our websites and games that have dynamic in-game advertising, you signify your assent to SCEA's privacy policy.
3. You may register or enhance your profile by linking your Facebook or Google accounts on NYTimes.com.
4. Sharing Your Information with Other Companies
5. You can delete cookies using your browser settings.
6. What Choices Do I Have?
7. You can visit our Web Sites without sharing personally identifiable information.

Annotation scheme/label choices:

OPP-115's annotation scheme consists of ten data practice categories:

1. *First party collection/use*: how and why a service provider collects user information.
2. *Third party sharing/collection*: how user information may be shared with or collected by third parties.
3. *User choice/control*: choices and control options available to users.
4. *User access, edit, & deletion*: if and how users may access, edit, or delete their information.
5. *Data retention*: how long user information is stored.
6. *Data security*: how user information is protected.

7. *Policy change*: if and how users will be informed about changes to the privacy policy.
8. *Do not track*: if and how Do Not Track signals for online tracking and advertising are honoured.
9. *International and specific audiences*: practices that pertain only to a specific group of users (e.g., children, Europeans, or California residents).
10. *Other*: additional sub-labels for introductory or general text, contact information, and practices not covered by the other categories.

We argue that revealing variations in understanding privacy policies is an underexplored area of research. This gap partly arises from the bias in using machine learning as the primary method and from the tendency to treat privacy as a uniform concept, overlooking cultural, educational, and gender differences. To address this, we conduct an empirical study using the OPP-115 corpus, which includes annotations from three annotators. We expand the characterization from three (individual, pairwise, and complete agreement) to seven: three individual, three pairwise, and one gold standard with full agreement. Unlike majority voting, pairwise agreement requires two specific annotators to agree. Complete agreement, while yielding fewer instances in the target class, may offer higher-quality annotations. Our study provides tangible measures for understanding how annotation variations impact machine learning performance. Rather than simply optimising metrics, we aim to reflect human disagreements through realistic reporting. We highlight two approaches to constructing gold standards—one ignoring disagreement and one considering it—without advocating for either, acknowledging that practical constraints often dictate these choices. We emphasise that disagreements in interpreting privacy statements are common, even among experts, and recognizing these differences is essential for user-centred privacy research.

Related work

High-quality labelled data are essential for supervised machine learning. Some argue that multiple annotators reduce human bias in evaluation (Artstein, 2017), but the number of annotators needed for high-quality annotation is unclear and often limited by budget (Mysore Sathyendra et al., 2017). Inter-rater reliability (IRR), or annotator agreement, is commonly used to measure annotation quality (Moallem, 2018). Typically, studies use either text annotated by any rater (union) or by the majority (majority vote), ignoring genuine annotator differences.

The assumption that annotators are interchangeable is also flawed, as differences exist between annotator populations and individuals (Hershcovich et al., 2022). Agreement studies show that variation can arise from heterogeneous data, complex labels, and annotator differences. Global agreement coefficients may mask these variations, while more detailed studies provide insights (Stevens et al., 2020).

A review of legal machine learning datasets found that disagreement during annotation is typically removed, with final corpora containing only '*gold standard*' annotations. Common strategies like majority vote, forced agreement, expert review, or arbitration do not account for disagreements (Prabhakaran et al., 2021). Issues of interpretation and non-transparency in reporting machine learning results, especially with legal documents, have also been flagged (Plank, 2022). Accurate predictions can build trust, but reproducibility depends on dataset validity and the discovery process (Herbert et al., 2023). Thus, documenting data preparation processes and their impact on model performance is crucial for trust and transparency in machine learning results (Bai et al., 2022).

Method

Towards this goal of using machine learning to provide privacy policy statement analysis that documents and considers annotation disagreements, our study provides ways to demonstrate how

to directly measure the impact of human (dis)agreement on machine learning model performance, by responding to the following research questions:

> RQ1: to what extent does reaching consensus amongst annotators impact the classification performance of traditional machine learning and deep learning models, respectively?
>
> RQ2: how do some alternative strategies used to create gold standards compare with the typical union and majority vote strategy?
>
> RQ3: what metrics can be used to better manage the trade-off between more annotated texts and more annotations for the same text?

To address these research questions, we conducted experiments using the OPP-115 corpora (Anaraky et al, 2019) with annotated text prepared through majority vote, union methods, and alternative strategies (individual, pairwise, and complete agreement). We tested how these strategies impact model performance using two traditional supervised learning algorithms—support vector machines (SVM) and Naïve Bayes (NB)—and two deep learning models, bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT).

Dataset

The OPP-115 Corpus comprises 13,209 sentences and the number of sentences in each target class are highly imbalanced (as shown in Figure 1). For example, there are only 6, 39, and 63 sentences with complete agreement in the Do Not Track, Data retention, and Policy Change categories respectively. The OPP-115 corpora were prepared by ten independent annotators who coded privacy policy text segments using several predefined categories. Each privacy segment was annotated by three independent annotators. We arranged the unique annotator IDs in each statement in ascending order and replaced them with annotators A, B, and C respectively. By our proposed alternative strategies, the gold standards used individual annotators (A or B or C), where 2 annotators reached consensus (A and B, or A and C or B and C) and where there was complete agreement (A and B and C agree on the annotations). We also replicated the strategy used in the original work where a text is deemed relevant when at least 2 out of 3 annotators (majority vote) reached agreement, or all the 3 annotators (union).

	Orig.	Independent			Pairwise			Complete	Majority	Union
Category	Kappa	A	B	C	AB	AC	BC	ABC		
Do not track	0.91	22	27	20	10	11	11	6	20	43
International and specific audiences	0.87	643	556	484	482	406	367	338	588	778
First party collection/use	0.76	2249	2297	2297	1512	1579	1575	1266	2156	3489
Third party sharing/collection	0.76	1614	1702	1646	1100	1098	1109	873	1600	2590
User access, edit and deletion	0.74	198	215	226	129	132	134	109	178	355
Policy change	0.73	189	160	185	84	99	91	63	150	326
Data security	0.67	291	373	289	175	163	186	128	271	563
User choice/control	0.61	580	649	619	307	300	325	224	485	1154
Data retention	0.55	105	118	135	48	61	55	39	86	233
Other	0.49	2183	1393	1863	1013	1248	882	758	1646	3087

Figure 1. Number of sentences in each gold standard

Annotations created in the original OPP-115 data were segmented text into paragraphs (Anaraky et al, 2019). In contrast, the unit of analysis in our experiments is a sentence, thus both the original text and the annotations were converted into sentences using version 4.1.2 of LingPipe

(Alabduljabbar et al., 2021) and the index position of each sentence was maintained and subsequently aligned with the index position of the manual annotations. Sentences were pre-processed using the NLTK Python package (Prabhakaran et al., 2021), including: converting words to lowercases, and removing punctuation and stop words. Terms appearing infrequently (less than 5 sentences) and very frequently (more than 95% of the sentences) were removed because their presence would contribute little to the classification performance (Amos et al., 2021). Annotation categories are not mutually exclusive, a sentence can be annotated as belonging to more than one multiple categories. Lastly, we framed the problem as a binary text classification task for each of the annotation categories.

Text classification

The classification experiments used two algorithm families: traditional models (SVM, NB) and deep learning models (BiLSTM, BERT). Ten-fold cross-validation evaluated each model, splitting the dataset into ten equal parts, with nine used for training and one for testing. The data was stratified to maintain a balanced proportion of positive and negative labels in each fold. Model performance was measured using standard metrics: precision, recall, F-1, and accuracy (Hamdani et al., 2021). For traditional models, feature selection was crucial. We used version 1.0.2 of Scikit-learn (Gordon, et al., 2022) and entropy-based selection, calculating information gain to choose the top 2000 features (Amos et al., 2021; Pedregosa et al., 2011). These features were then used to construct the sentence-term matrix for the test set. TF-IDF was considered but not used due to its inability to account for target class distribution (Srinath et al., 2021).

Results

We report our findings regarding how much annotator agreement impacts the performance of automated approaches.

Alternative methods: independent, pairwise, complete agreement

We found that increasing the level of agreement from independent to pairwise to complete improved F-1, accuracy, precision, and recall across nearly all categories and classifiers (Figure 2). Complete agreement yielded the best F1 scores for several categories, including First Party Collection/Use and Third-Party Sharing/Collection. For Third Party Sharing/Collection, precision improved by 5% and recall by 4%. First Party Collection/Use saw a 7% improvement in precision and 8% in recall. The '*Other*' category showed significant gains, with F1 improving from 0.82 to 0.91.

In some categories, like User Choice/Control and International and Specific Audiences, pairwise agreement performed as well as complete agreement, with recall and accuracy reaching 0.97. Categories with fewer examples, such as Data Retention, had high metrics but raised concerns about generalizability.

Certain classifiers handled data inconsistencies better. For Third Party Sharing/Collection, precision improved from 0.85 (independent) to 0.91 (complete agreement) across all classifiers. The difference in model performance was minor (within 0.02), but the impact of different gold standards was more significant, ranging from 0.03 to 0.06 (Figures 2 and 3). Third Party Sharing/Collection also showed notable improvements in F-1, precision, recall, and accuracy, with precision and F1 for First Party Collection/Use improving by at least 0.07.

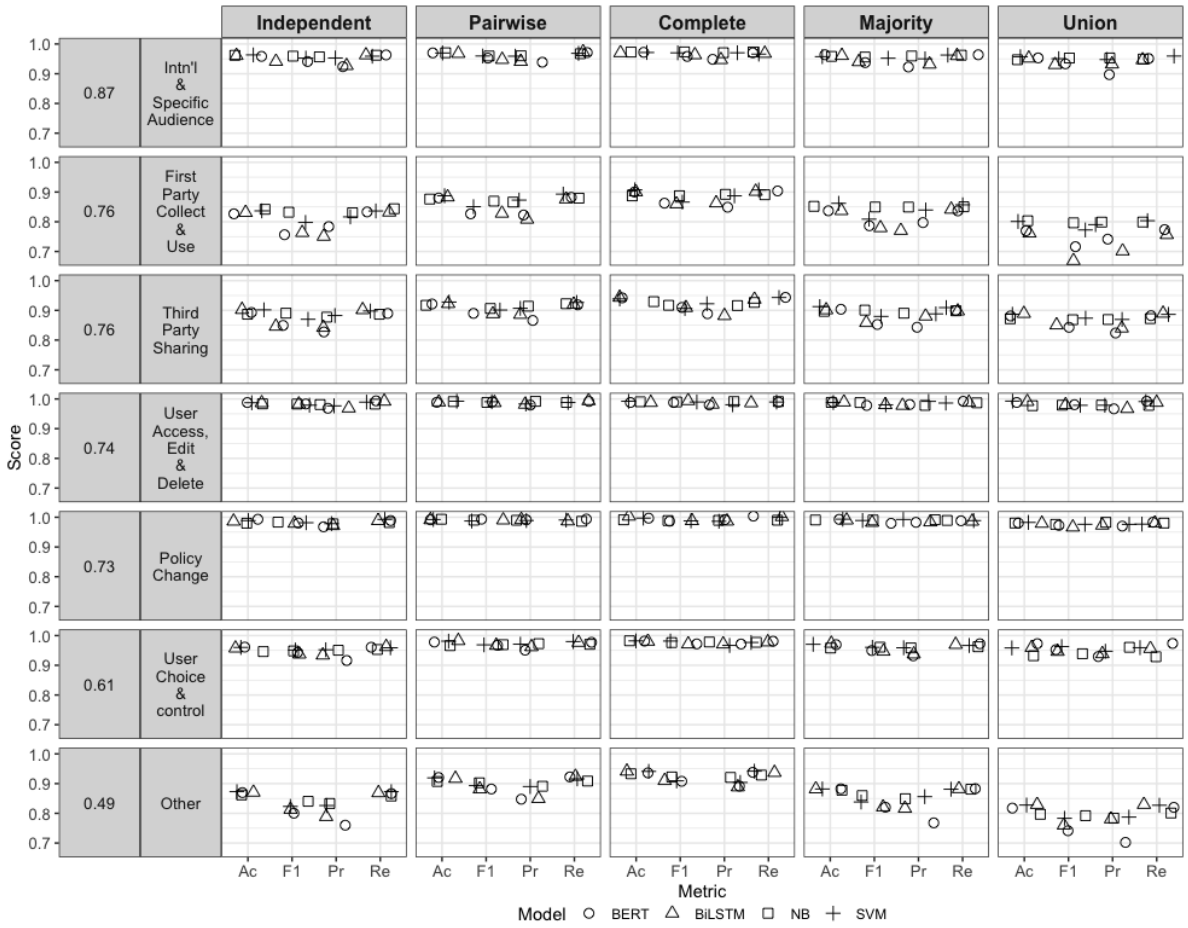


Figure 2. Average model performance for increasing levels of consensus (independent, pairwise, and complete agreement) and increasing levels of disagreement (majority vote, union) gold standards *(Ac=accuracy, F1, Pr=precision, and Re=recall)

Category	Sents	Model	Accuracy			Recall			Precision			F1		
			Complete	Union	Δ	Complete	Union	Δ	Complete	Union	Δ	Complete	Union	Δ
International and specific audiences		BERT	0.97	0.95	0.02	0.97	0.95	0.02	0.95	0.90	0.05	0.96	0.93	0.03
	Complete 338	BiLSTM	0.97	0.95	0.02	0.97	0.95	0.02	0.95	0.93	0.02	0.96	0.93	0.03
	Union 778	NB	0.97	0.95	0.02	0.97	0.95	0.02	0.97	0.95	0.02	0.97	0.95	0.02
		SVM	0.97	0.96	0.01	0.97	0.96	0.01	0.97	0.95	0.02	0.97	0.95	0.02
First party collection/use		BERT	0.90	0.77	0.13	0.90	0.77	0.13	0.85	0.74	0.11	0.86	0.72	0.14
	Complete 1266	BiLSTM	0.90	0.76	0.14	0.90	0.76	0.14	0.86	0.70	0.16	0.86	0.67	0.19
	Union 3489	NB	0.89	0.80	0.09	0.89	0.80	0.09	0.89	0.80	0.09	0.89	0.80	0.09
		SVM	0.91	0.80	0.11	0.91	0.80	0.11	0.89	0.79	0.10	0.87	0.77	0.10
Third party sharing/collection		BERT	0.94	0.88	0.06	0.94	0.88	0.06	0.89	0.82	0.07	0.91	0.84	0.07
	Complete 873	BiLSTM	0.94	0.89	0.05	0.94	0.89	0.05	0.88	0.84	0.04	0.91	0.85	0.06
	Union 2590	NB	0.93	0.87	0.06	0.93	0.87	0.06	0.92	0.87	0.05	0.92	0.87	0.05
		SVM	0.94	0.89	0.05	0.94	0.89	0.05	0.92	0.87	0.05	0.91	0.87	0.04
User access, edit and deletion		BERT	0.99	0.99	0.00	0.99	0.99	0.00	0.98	0.97	0.01	0.99	0.98	0.01
	Complete 109	BiLSTM	0.99	0.99	0.00	0.99	0.99	0.00	0.98	0.97	0.01	0.99	0.98	0.01
	Union 355	NB	0.99	0.98	0.01	0.99	0.98	0.01	0.99	0.98	0.01	0.99	0.98	0.01
		SVM	0.99	0.99	0.00	0.99	0.99	0.00	0.98	0.98	0.00	0.99	0.98	0.01
Policy change		BERT	1.00	0.98	0.02	1.00	0.98	0.02	0.99	0.97	0.02	0.99	0.97	0.02
	Complete 63	BiLSTM	1.00	0.98	0.02	1.00	0.98	0.02	0.99	0.97	0.02	0.99	0.97	0.02
	Union 326	NB	0.99	0.98	0.01	0.99	0.98	0.01	0.99	0.98	0.01	0.99	0.98	0.01
		SVM	1.00	0.98	0.02	1.00	0.98	0.02	0.99	0.98	0.01	0.99	0.98	0.01
User choice/control		BERT	0.98	0.97	0.01	0.98	0.97	0.01	0.97	0.93	0.04	0.97	0.95	0.02
	Complete 224	BiLSTM	0.98	0.96	0.02	0.98	0.96	0.02	0.97	0.94	0.03	0.97	0.95	0.02
	Union 1154	NB	0.98	0.93	0.05	0.98	0.93	0.05	0.98	0.96	0.02	0.98	0.94	0.04
		SVM	0.98	0.96	0.02	0.98	0.96	0.02	0.97	0.95	0.02	0.98	0.96	0.02
Other		BERT	0.94	0.82	0.12	0.94	0.82	0.12	0.89	0.70	0.19	0.91	0.74	0.17
	Complete 758	BiLSTM	0.94	0.83	0.11	0.94	0.83	0.11	0.89	0.78	0.11	0.91	0.76	0.15
	Union 3087	NB	0.93	0.80	0.13	0.93	0.80	0.13	0.92	0.78	0.14	0.92	0.79	0.13
		SVM	0.94	0.83	0.11	0.94	0.83	0.11	0.90	0.79	0.11	0.91	0.78	0.13

Figure 3. Average model performance for complete and union gold standards

Typical gold standard methods: union and complete

In general, we also found increasing the level of agreement improves model performance: the majority vote results outperform the union across all metrics and all classifiers, in all categories. Figure 4 shows the original Fleiss' Kappa statistic against the F1 score for complete and union gold standards. The range of Kappa in the OPP-115 collection ranges from moderate to very good, so it is possible that larger variations might show a correlation with F1, but these results suggest that: (a) Kappa is not a good substitute for F1 scores produced using different gold standards; (b) the difference between the standard metrics produced from complete and union experiments might provide a more realistic way to convey the impact of disagreement.

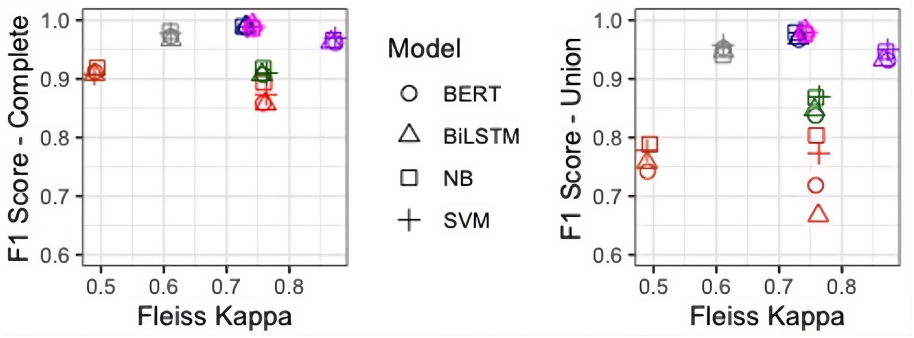


Figure 4. Fleiss Kappa versus F1-Score for complete and union gold standards

Discussion

Common methods for reporting machine learning results often focus on the model and its tuning parameters, neglecting variations in human-labelled training data. In privacy documentation,

diverse terminology and interpretations make it essential to consider human discrepancies in machine learning analysis. Traditional metrics don't always reflect the Fleiss Kappa statistic or the impact of human consensus. Based on our research, we recommend:

1. **Use multiple annotators:** for at least a subset of the corpus, multiple annotators should be employed to capture human variability. Reducing the number of categories per annotator could enhance annotation quality and efficiency.
2. **Iterative annotation:** apply an iterative approach, directing annotators to categories with less agreement. This can optimise human effort and maintain annotation quality, using complete and union gold standards as metrics.
3. **Differentiate gold standards:** use the difference between complete and union gold standards to assess how well machine learning results align with human judgments. This can provide insight into the model's accuracy in mirroring human interpretations.
4. **Avoid aggregation:** report results separately for each category rather than aggregating them. Specific categories may have varying expectations, and detailed performance information is crucial for understanding model effectiveness in real-world applications.

Conclusion

Privacy policy statements are vital for regulatory compliance and user data decisions, yet many are unreadable and often ignored. Machine learning could help by automating information extraction, but current reporting practices that don't align with human judgement undermine trust. Unlike previous methods using Fleiss' Kappa to measure disagreement, we propose an approach that uses independent, pairwise, and complete agreement in gold standards. We acknowledge that our data set is limited as it used only the OPP-115 corpus which may limit the generalizability of this study. Given this limitation, our preliminary results show that higher agreement improves precision, recall, F1, and accuracy, while more disagreement reduces these metrics.

Disagreements in privacy statement interpretation are more complex than fact-based tasks, and inter-rater reliability alone may not suffice to measure model performance. Traditional metrics like Cohen's or Fleiss Kappa are inadequate for skewed data. We suggest using precision, recall, F1, and accuracy to evaluate how different gold standards affect performance, which is crucial given the evolving nature of privacy content.

With new collections of privacy statements surpassing a million entries (Bannihatti Kumar et al., 2020; Thorleiksdóttir et al., 2022), investing in annotation adjudication is urgent. Quality assurance involves decisions such as storing multiple annotators' data and measuring their agreement (Mousavi Nejad et al., 2020). Our study highlights the need for multiple annotators on subsets of texts to assess the impact of human judgement on metrics. This may conflict with current practices aimed at maximising annotated data but is essential for realistic metric representation. An iterative approach can help allocate resources effectively, and text classification results should be reported by category rather than aggregated.

Our study is the first to explicitly raise the question of disagreements of annotating privacy policies documents and highlight the value and significance of studying such disagreements. How exactly disagreement in understanding and/or interpreting privacy policies can be leveraged remains unexplored, and future research needs to understand where and why disagreements occur when it comes to privacy policy interpretation, as there can be potentially disagreements for completely different reasons and hence require different treatment or solution. For example, there may be disagreement that originates from a lack of knowledge, linguistic ambiguity, or underlying differences in preference, each would require completely different solutions.

When marking up raw text, annotators need the flexibility to decide the appropriate text boundaries that capture the target category. Before the initial annotations can be used to construct a classifier, the unit of analysis, such as a paragraph, sentence (used in this analysis) or some other predefined '*span*' must be established. This choice impacts the predictive performance of any model constructed, more work is needed to establish what span is optimal for a given task and to quantify the impact of this decision. We have introduced a new performance metric – the difference between complete and union gold standards - that directly measures the impact of human agreement using the same metrics that are commonly used to measure an automated system. However, situated empirical user studies are needed to establish if this new metric is successful in making machine-learning models more transparent.

Acknowledgements

We would like to thank all the reviewers for the feedback on this paper. There is no funding for this research to report.

About the authors

Tian Wang is Postdoctoral Associate at the CyLab Security and Privacy Institute of Carnegie Mellon University. She received their Ph.D. from University of Illinois at Urbana Champaign, and her research interests are in mobile and app security and privacy. She can be contacted at tianwan2@andrew.cmu.edu

Yuanye Ma is Senior Research Associate at University of Illinois Discovery Partners Institute. She received their Ph.D. from University of North Carolina at Chapel Hill, and her research interests are in user-centered privacy, natural language processing and information ethics. She can be contacted at yuanyem@uillinois.edu

Catherine Blake is Professor and Associate Dean for Academic Affairs in School of Information Sciences, University of Illinois at Urbana Champaign. She received her PhD from University of California, Irvine. Her research interests are in biomedical informatics, natural language processing, evidence-based discovery, learning health systems, socio-technical systems, data analytics, literature-based discovery. She can be contacted at clblake@illinois.edu

Masooda Bashir is Associate Professor in School of Information Sciences, University of Illinois at Urbana Champaign. She received their Ph.D. from Purdue University, and her research interests are in the interface of information technology, human psychology, and society; especially how privacy, security, and trust intersect from a psychological point of view with information systems. She can be contacted at mnb@illinois.edu

Ryan Wang is a PhD student in the School of Information Sciences at the University of Illinois Urbana-Champaign. He is interested in natural language processing, machine learning, and bioinformatics. He can be reached at hywang3@illinois.edu.

References

Alabduljabbar, A., Abusnaina, A., Meteriz-Yildiran, Ü., & Mohaisen, D. (2021). TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights. Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society, 103–118. https://doi.org/10.1145/3463676.3485608

Anaraky, R. G., Cherry, D., Jarrell, M., & Knijnenburg, B. (2019). Testing a comic-based privacy policy. In *The 15th Symp. on Usable Privacy and Security*.

Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation*, 297-313.

Azhagusundari, B., & Thanamani, A. S. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering* (IJITEE), 2(2), 18-21.

Bai, F., Ritter, A., & Xu, W. (2021). Pre-train or annotate? domain adaptation with a constrained budget. *arXiv preprint arXiv*:2109.04711.

Bannihatti Kumar, V., Iyengar, R., Nisal, N., Feng, Y., Habib, H., Story, P., Cherivirala, S., Hagan, M., Cranor, L., Wilson, S., Schaub, F., & Sadeh, N. (2020). Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. *Proceedings of The Web Conference 2020*, 1943–1954. https://doi.org/10.1145/3366423.3380262

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analysing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Braun, D. (2024). I beg to differ: How disagreement is handled in the annotation of legal machine learning data sets. *Artificial Intelligence and Law*, 32(3), 839–862. https://doi.org/10.1007/s10506-023-09369-4

Chen, R., Fang, F., Norton, T., McDonald, A. M., & Sadeh, N. (2021). Fighting the Fog: Evaluating the Clarity of Privacy Disclosures in the Age of CCPA. *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, 73–102. https://doi.org/10.1145/3463676.3485601

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., & Bernstein, M. S. (2022, April). Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.

Grosman, J. S., Furtado, P. H. T., Rodrigues, A. M. B., Schardong, G. G., Barbosa, S. D. J., & Lopes, H. C. V. (2020). Eras: Improving the quality control in the annotation process for Natural Language Processing tasks. *Information Systems*, 93, 101553. https://doi.org/10.1016/j.is.2020.101553

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34.

Hamdani, R. E., Mustapha, M., Amariles, D. R., Troussel, A., Meeùs, S., & Krasnashchok, K. (2021). A combined rule-based and machine learning approach for automated GDPR compliance checking. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 40–49. https://doi.org/10.1145/3462757.3466081

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. *Proceedings of the 27th USENIX Conference on Security Symposium*, 531–548.

Herbert, F., Becker, S., Schaewitz, L., Hielscher, J., Kowalewski, M., Sasse, A., Acar, Y., & Dürmuth, M. (2023). A World Full of Privacy and Security (Mis)conceptions? Findings of a Representative Survey in 12 Countries. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–23. https://doi.org/10.1145/3544548.3581410

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., & Søgaard, A. (2022).

Challenges and Strategies in Cross-Cultural NLP. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6997–7013). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.482

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2), 1.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Li, J., Liu, L., Le, T. D., & Liu, J. (2020). Accurate data-driven prediction does not mean high reproducibility. *Nature Machine Intelligence*, 2(1), 13–15. https://doi.org/10.1038/s42256-019-0140-2

LingPipe Alias-i. (2008). 4.1. 0. URL http://alias-i.com/lingpipe (2008).

Liu, F., Wilson, S., Schaub, F., & Sadeh, N. (2016). Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In 2016 AAAI Fall Symposium Series.

Moallem, A. (2018). Do You Really Trust "Privacy Policy" or "Terms of Use" Agreements Without Reading Them? In D. Nicholson (Ed.), *Advances in Human Factors in Cybersecurity* (pp. 290–295). Springer International Publishing. https://doi.org/10.1007/978-3-319-60585-2_27

Mousavi Nejad, N., Jabat, P., Nedelchev, R., Scerri, S., & Graux, D. (2020). Establishing a Strong Baseline for Privacy Policy Classification. In M. Hölbl, K. Rannenberg, & T. Welzer (Eds.), *ICT Systems Security and Privacy Protection* (pp. 370–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-58201-2_25

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in python journal of machine learning research. *Journal of machine learning research*, 12, 2825-2830.

Pepperberg, I. M. (1988). An interactive modeling technique for acquisition of communication skills: Separation of "labeling" and "requesting" in a psittacine subject. *Applied Psycholinguistics*, 9(1), 59–76. https://doi.org/10.1017/S014271640000045X

Plank, B. (2022). *The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation* (arXiv:2211.02570). arXiv. https://doi.org/10.48550/arXiv.2211.02570

Prabhakaran, V., Davani, A. M., & Díaz, M. (2021). *On Releasing Annotator-Level Labels and Information in Datasets* (arXiv:2110.05699). arXiv. https://doi.org/10.48550/arXiv.2110.05699

Reidenberg, J. R., Breaux, T., Cranor, L. F., French, B., Grannis, A., Graves, J., Liu, F., McDonald, A., Norton, T., Ramanath, R., Russell, N. C., Sadeh, N., & Schaub, F. (2014). *Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding* (SSRN Scholarly Paper 2418297). https://doi.org/10.2139/ssrn.2418297

Mysore Sathyendra, K., Wilson, S., Schaub, F., Zimmeck, S., & Sadeh, N. (2017). Identifying the Provision of Choices in Privacy Policy Text. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2774–2779). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1294

Srinath, M., Wilson, S., & Giles, C. L. (2021). Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6829–6839. https://doi.org/10.18653/v1/2021.acl-long.532

Stevens, L. M., Mortazavi, B. J., Deo, R. C., Curtis, L., & Kao, D. P. (2020). Recommendations for Reporting Machine Learning Analyses in Clinical Research. *Circulation. Cardiovascular Quality and Outcomes*, *13*(10), e006556. https://doi.org/10.1161/CIRCOUTCOMES.120.006556

Tang, J., Shoemaker, H., Lerner, A., & Birrell, E. (2021). Defining privacy: How users interpret technical terms in privacy policies. *Proceedings on Privacy Enhancing Technologies*.

Thorleiksdóttir, T., Renggli, C., Hollenstein, N., & Zhang, C. (2022). Dynamic Human Evaluation for Relative Model Comparisons. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5946–5955). European Language Resources Association. https://aclanthology.org/2022.lrec-1.639

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Giovanni Leon, P., Schaarup Andersen, M., Zimmeck, S., Sathyendra, K. M., Russell, N. C., Norton, T. B., Hovy, E., Reidenberg, J., & Sadeh, N. (2016). The Creation and Analysis of a Website Privacy Policy Corpus. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1330–1340). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1126

Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, No. 412-420, p. 35).

Zaeem, R. N., German, R. L., & Barber, K. S. (2018). PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Trans. Internet Technol.*, *18*(4), 53:1-53:18. https://doi.org/10.1145/3127519

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

A more advanced group polarization measurement approach based on LLM-based agents and graphs

Zixin Liu, Ji Zhang, and Yiran Ding

DOI: https://doi.org/10.47989/ir30iConf47578

Abstract

Introduction. Group polarization is an important research direction in social media content analysis, attracting many researchers to explore this field. Therefore, how to effectively measure group polarization has become a critical topic. Measuring group polarization on social media presents several challenges that have not yet been addressed by existing solutions. First, social media group polarization measurement involves processing vast amounts of text, which poses a significant challenge for information extraction. Second, social media texts often contain hard-to-understand content, including sarcasm, emojis, and internet slang. Additionally, group polarization research focuses on holistic analysis, while texts is typically fragmented. These challenges indicate that a new solution needs to be proposed.

Method. To address these challenges, we designed a solution based on a multi-agent system and used a graph-structured community sentiment network (CSN) to represent polarization states. Furthermore, we developed a metric called community opposition index (COI) based on the CSN to quantify polarization.

Conclusion. We tested our multi-agent system through a zero-shot stance detection task and achieved outstanding results, which proved its significant value in terms of usability and accuracy.

Introduction

With the development of internet technology, social media has gained widespread popularity. *GLOBAL DIGITAL REPORT* (https://datareportal.com/global-digital-overview) indicates that platforms such as Facebook, YouTube, and TikTok boast billions of users worldwide. Social media has become a key avenue for the public to express opinions and engage in discussions. Its anonymity and convenience enable users to freely express their true views, thereby shaping social media public opinion. As a result, research in this field has also flourished.

In these studies, research from the perspective of group polarization holds a significant position. The concept of group polarization was first introduced by Stoner, who observed that group decisions tend to be more extreme compared to individual decisions (Stoner, 1961; Isenberg, 1986). In the internet era, group polarization is broadly defined as the divergence of public opinions or stances. Building on this definition, researchers have conducted extensive and comprehensive studies on various issues related to group polarization. One of the fundamental research problems in the field of social media group polarization is its measurement. Early measurement methods based on statistical approaches suffered from issues such as overly simplistic for the complexity of social media dynamics (Bilal et al., 2019; Gaurav et al., 2013; Hart et al., 2020; Jaidka et al., 2018; Tumasjan et al., 2010). Current mainstream methods, such as text clustering or sentiment classification (Belcastro et al., 2020; Jiang et al., 2018; Ribeiro et al., 2017; Tyagi et al., 2020), struggle to balance efficiency and interpretability. While some researchers have made significant progress by focusing on the relationships between different viewpoints, these studies still fall short in understanding the deeper nuances of opinion stances and their evolution (Boxell et al., 2020; Iyengar et al., 2019; Jiang et al., 2023; Lelkes et al., 2024; Maia et al., 2023).

To address the existing issues in measuring group polarization and improve efficiency, accuracy, and interpretability, we propose a new group polarization measurement approach based on LLM-based agents and graphs. This approach draws inspiration from the stance detection task in natural language processing (NLP) and the earlier *sentiment thermometer* method. We use a community sentiment network (CSN) represented by a graph structure to model the polarization state, where LLM-based agents are employed to construct the network. Additionally, we design polarization measurement metrics based on CSN. To validate the effectiveness of our approach, we tested the module responsible for constructing CSN on zero-shot stance detection tasks, and the results demonstrated its superiority in capturing the nuances of group polarization.

In summary, our contributions are as follows: (1) we propose a temporal community sentiment network (CSN) to represent the polarization state over time. (2) we introduce LLM-based agents for stance detection into group polarization measurement, significantly enhancing both efficiency and accuracy. (3) we propose a more robust metric based on the CSN, community opposition index (COI).

Related work

Opinion polarization measurement

As research on group polarization deepens, extensive exploration has also been conducted on the measurement of group polarization, leading to the development of a relatively comprehensive system of measurement approaches. Existing research suggests that the current mainstream group polarization measurement schemes can be primarily divided into three categories: volume-based, sentiment-based, and network-based (Bilal et al., 2019; Jaidka et al., 2018). We will discuss the characteristics of these three measurement schemes and their shortcomings when applied to the task of measuring group polarization on social media.

Volume-based approaches primarily rely on statistical methods and were widely applied in the early research of group polarization. Early researchers collected data through surveys and

experiments and used statistical analysis to obtain relevant polarization results. In current trend of exploring group polarization via social media, volume-based schemes focus more on various data metrics and employ statistical methods in research. Notable examples include Gaurav et al.'s political polarization study based on the moving average aggregate probability method, Tumasjan et al.'s analysis of political polarization using the LIWC tool, and Hart et al.'s use of multidimensional statistical analysis to analyse polarization during COVID-19 (Gaurav et al., 2013; Tumasjan et al., 2010; Hart et al., 2020).

However, existing studies suggest that volume-based schemes have limitations in terms of capturing information and analysing large datasets. They fail to accurately understand the opinions and sentiments conveyed in the text and generally rely on broad statistical metrics (e.g., word frequency, likes, bookmarks, etc) to gather limited information. The lack of rapid information capture and in-depth understanding makes these techniques less effective for tracking and real-time analysis of group polarization, and they also fall short in terms of precision in measuring polarization.

Compared to volume-based approaches, sentiment-based approaches place greater emphasis on the meaning and emotions conveyed in the text. Typically, sentiment-based approaches are grounded in natural language processing (NLP) and analyse social media text from the perspectives of opinions and emotions. These methods generally follow two main strategies. The first involves clustering texts based on the similarity of sentiments, such as the IOM-NN method proposed by Belcastro for accurately detecting emotional information in political polarization (Belcastro et al., 2020). The second strategy leverages deep learning for direct sentiment classification, exemplified by Tyagi et al.'s research on polarization driven by climate change, and the explorations by Ribeiro et al. and Jiang et al. on the relationship between misinformation and polarization (Tyagi et al., 2020; Ribeiro et al., 2017; Jiang et al., 2018).

Unfortunately, both strategies face notable challenges in practice. For text clustering, current clustering algorithms are relatively coarse and simplistic, making it difficult to distinguish between disruptive information (such as advertisements or neutral statements) and significant content. Moreover, they do not account for the specific relationships between subgroups or their contribution to polarization, resulting in outcomes that lack precision and interpretability. In sentiment classification, current methods often rely on binary classifications, failing to capture the intensity of emotions. This oversimplified method negatively impacts both the interpretability and accuracy of polarization measurement.

It is also worth noting that in political polarization research, a method called the *sentiment thermometer* has been widely adopted (Boxell et al., 2020; Iyengar et al., 2012; Iyengar et al., 2019; Lelkes et al., 2017; Wakefield et al., 2023). This approach uses surveys to gather voters' emotional scores toward various political parties, thus enhancing the precision and interpretability of the analysis. However, this method is costly, limited by small sample sizes, and not well-suited for measuring polarization in the context of social media, as it generally considers only two opposing groups, whereas social media often involves more complex, multi-group dynamics (Iyengar et al., 2012).

Network-based approaches represent an approach that evaluates group polarization by considering social positions and relationships among groups, focusing on emotional direction and the stance between subgroups to better measure polarization levels. Traditional social network analysis in group polarization studies explores peripheral connections around core opinions to assess subgroup positions and emotions (Bravo et al., 2015; Conover et al., 2011; Garcia et al., 2015; Guerra et al., 2013; Medaglia et al., 2017; Vicario et al., 2017). Some researchers have further developed this by dynamically simulating the process of group polarization to explore its evolutionary pathways (Maia et al., 2023; Santos et al., 2021). With the advancement of graph neural

networks (GNNs), network-based approaches have been enhanced, as demonstrated by valuable explorations from researchers such as Xiao et al., Zhang et al., and Jiang et al., who utilized sentiment networks and GNNs in their studies (Xiao et al., 2020; Zhang et al., 2019; Jiang et al., 2023). While these studies have achieved promising results in improving the scientific rigor and interpretability of group polarization research, they generally lack detailed analysis of the textual content. Moreover, social division within purely social networks does not necessarily result from group polarization, raising questions about the accuracy of some conclusions drawn from these methods.

LLM-based agents

With the introduction of OpenAI's GPT series of large language models (LLMs), numerous research fields have incorporated or examined GPT's capabilities (Brown et al., 2020). In the realm of group polarization research, some researchers have also explored its applications. For instance, Lu et al. used agent-based simulations to model group polarization dynamics, while Zhang et al. employed LLM-Based Agents to detect stances within polarized groups (Lu, 2024; Zhang, 2022). These studies have yielded promising results, demonstrating the feasibility and value of applying large language models in this field.

Method

As we mentioned in the first section, to achieve an accurate measurement of group polarization, our proposed method consists of three parts, specifically: (1) a community sentiment network (CSN) used to represent subgroups, the emotions between and within subgroups. (2) an efficient multi-agent system for CSN construction. (3) a group polarization metric based on the CSN, community opposition index (COI).

The set of opinion texts will be used to identify subgroups and conduct sentiment analysis through the multi-agent system, forming a CSN. The current polarization measurement result for the time slice can then be calculated using the COI.

Community sentiment network

The community sentiment network (CSN) is an extension of the *sentiment thermometer* method. The traditional *sentiment thermometer* could only be applied to two subgroups (Iyengar, 2012). CSN extends the *sentiment thermometer* to a directed cyclic graph that involves emotions between multiple subgroups (see Figure 1). Let $G = (V, E)$ be a graph, where V is the vertex set containing subgroups and E is the set of edges representing the sentimental relationships between subgroups. Each edge $e \in E$ is defined as (u, v, s) where $u, v \in V$ and s is the sentiment score. It should be noted that nodes v and u are allowed to be the same node, meaning self-loops are permitted. The score s can be either positive or negative, reflecting the positive or negative nature of the sentiment.

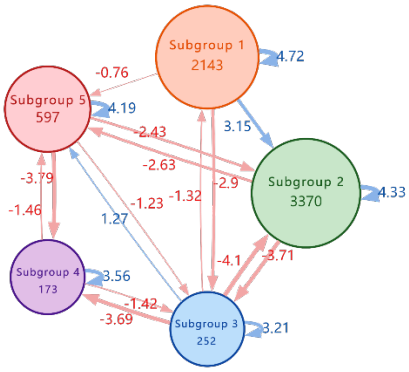


Figure 1. An example of a CSN generated by Graphviz based on comments scraped from Weibo under topics related to the Russia-Ukraine conflict, within a random time window

Unlike traditional social networks, CSN uses sentiment rather than interactions as the basis for constructing connections. Also, CSN clearly illustrates the various subgroups with different stances within the target time period and reveals the emotional states between subgroups as well as the internal cohesion of each subgroup. Therefore, compared to clustering results or social networks, CSN significantly highlights the contributions of different subgroups to polarization, providing greater interpretability of the polarization state.

The construction of the CSN involves multiple issues, such as subgroup identification, stance detection of opinion information, and sentiment recognition. However, existing research in the field of group polarization is insufficient to provide effective solutions to these issues. Therefore, inspired by the advancements in stance detection tasks (Xiao et al., 2020), we designed a multi-agent system based on LLM-based agents (see Figure 2).

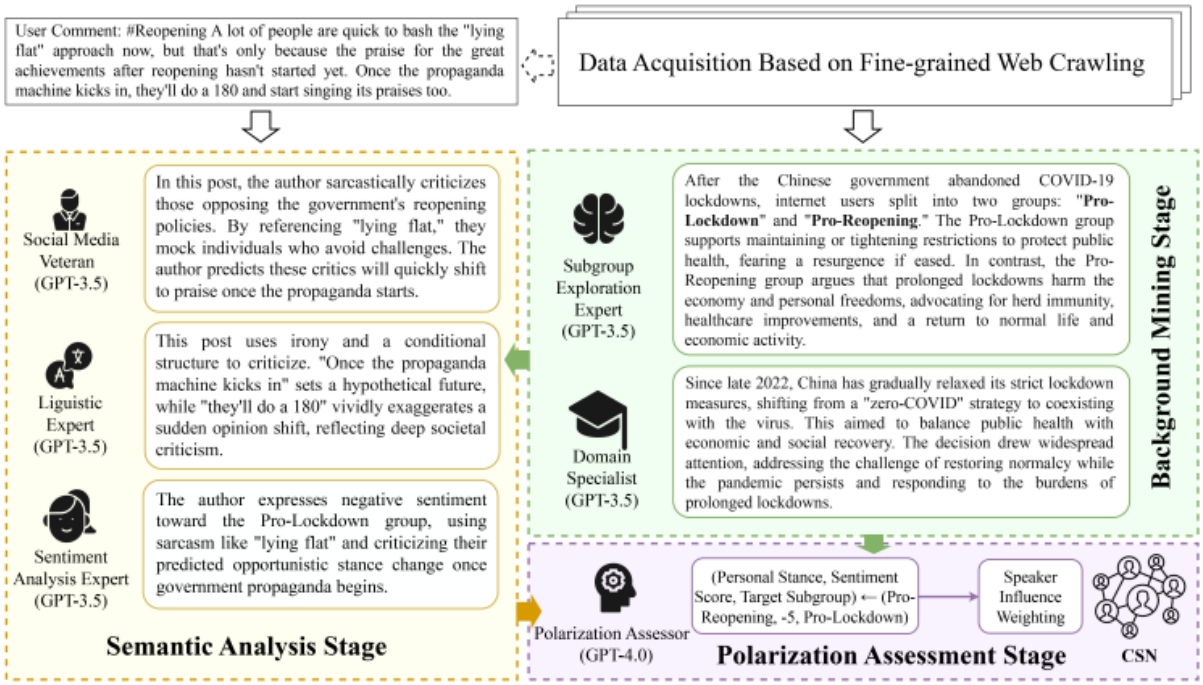


Figure 2. The structure of multi-agent system for CSN construction, containing background mining stage, semantic analysis stage and polarization assessment stage

In summary, our multi-agent system is composed of three stages: the background mining stage, the semantic analysis stage, and the polarization assessment stage. The background mining stage

consists of subgroup exploration expert and domain specialist, the semantic analysis stage includes social media veteran, linguistic expert, and sentiment analysis expert. The polarization assessment stage comprises polarization assessor. The agents collaborate through a series of interactions to provide accurate subgroup division results and reliable sentiment scores (see Figure 3). We will provide a detailed explanation of each stage and the functions of the respective agents in the following paragraphs.

```
Input: Weibo comments data within a specified time range
Output: Set of triplets as analysis results
procedure AnalyzeComments(comments)
    bg ← DomainExpert(comments)                    ▷ bg for background
    sg ← ∅                                         ▷ sg for subgroups
    uncertainComments ← []
    SubgroupMiningExpert(bg)
    for each comment in comments do
        grp ← SubgroupMiningExpert(comment)
        if not grp then
            uncertainComments.add(comment)
            if length(uncertainComments) reaches threshold then
                grp ← HumanExpertHelp(uncertainComments)
                sg.add(grp)
                uncertainComments ← []
            end if
        else
            sg.add(grp)
        end if
    end for
    Initialize Experts:
    SocialMediaVeteran(sg, bg)
    LinguisticExpert(bg)
    AffectAnalysisExpert(sg)
    PolarizationJudge(sg, bg)
    triplets ← ∅
    for each comment in comments do
        out1 ← SocialMediaVeteran(comment)
        out2 ← LinguisticExpert(comment)
        out3 ← AffectAnalysisExpert(comment, out1, out2)
        out4 ← PolarizationJudge(comment, out3)
        if out4.group then
            triplet ← (out4.group, out4.score, out4.targetGroup)
        else
            triplet ← (null, out4.score, out4.targetGroup)
        end if
        triplets.add(triplet)
    end for
    return triplets
end procedure
```

Figure 3. Construction of group affect network

Background mining stage

For the construction of the CSN, the first issue we need to address is how the multi-agent system understands the overall event within the context of the event. To solve this, we propose the Background Mining Stage, which uses textual information to understand the event and identify potential subgroups. Its functionality can be described as follows:

Input: all the comment texts related to the target topic (sampled if necessary).

Output: a description of the event's background, all potential subgroups present in the topic, and detailed descriptions of each subgroup.

Domain specialist. Domain specialist is primarily responsible for extracting the event background described in the comment texts. Their main tasks include exploring the core event of the topic and key stakeholders. The domain specialist develops a comprehensive description of the event's

timeline and related parties, providing this background information to the subgroup exploration expert and other subsequent stages of the process.

Subgroup exploration expert. The task of the subgroup exploration expert is to use the background information provided by the domain specialist along with the source texts to identify potential subgroups involved in the event and summarize the possible speaking patterns of each subgroup's members. This requires the expert to explore the organizations, stances, religions, and other social identities referenced in the texts and form subgroups based on the similarity of their expressions. It is important to note that if there is a significant amount of unclassifiable content, the expert is permitted to consult human experts for clarification.

Semantic analysis stage

Building on the background information, semantic analysis stage needs to address the primary challenge of accurately interpreting the emotions conveyed in the texts, especially when slang, homophones, sarcasm, and other nuanced expressions are present. To achieve better results on this complex task, we design the system with the specific attributes of social media in mind. Its functionality can be described as follows:

Input: comment texts under the target topic and results from background mining stage.

Output: sentiment analysis result.

Social media veteran. The social media veteran is one of the key agents responsible for semantic understanding of social media content. Its main role is to explore the patterns and characteristics of language expression on social media platforms. The agent needs to interpret the actual meanings of hashtags, internet slang, emojis, and other unique forms of expression commonly used on social media. After this analysis, the social media veteran passes the comprehended information to the sentiment analysis expert for further processing.

Linguistic expert. The linguistic expert is another key agent responsible for the semantic understanding of social media content. Unlike the social media veteran, the linguistic expert focuses on analysing the text from a linguistic perspective, examining aspects such as grammatical structure, rhetorical devices, word choice, and tense. The analysis results are also passed to the sentiment analysis expert. It is important to note that the linguistic expert's analysis is not conducted in isolation; the background information provided by the domain expert supports and informs the linguistic analysis.

Sentiment analysis expert. The sentiment analysis expert is the agent responsible for synthesizing various inputs to determine the final sentiment and its direction. It combines the emotional language present in the text with the semantic analysis results provided by other agents, such as the social media veteran and the linguistic expert, to derive the sentiment of the text. Additionally, it utilizes the subgroup information provided by the subgroup exploration expert to identify the potential target of the sentiment. The results of this sentiment analysis will serve as the output of the sentiment analysis stage and will be passed to the next stage.

Polarization assessment stage

The primary task of the polarization assessment stage is to utilize the information from the background mining stage and the semantic analysis stage to generate CSN in the form of triplets. Its functionality can be described as follows:

Input: all information from the background mining stage (including potential subgroups and event background), each text from the target topic, and their sentiment analysis results.

Output: sentiment expressed in the form of triplets for each comment: (personal stance, sentiment score, target subgroup).

Polarization assessor. The polarization assessor is the core agent of the polarization assessment stage. It is responsible for analysing the author's stance, sentiment score, and the target group of the sentiment for each comment, based on the information passed from the other stages. The polarization assessor integrates this information into triplets. With the multi-dimensional and in-depth analysis provided by the other stages, the polarization assessor can make precise judgments and give credible sentiment score.

CSN construction. To construct the final community sentiment network (CSN) based on the sentiment triplets from the existing comments, we designed the relevant algorithm (see Figure 4, Table 1 explains some variables). We use an adjacency matrix, adjMatrix, to represent the CSN, where adjMatrix[i][j] represents the sentiment score of subgroup i towards subgroup j. We first use all the triplets to build an initial network and then merge the nodes that belong to the same subgroup. During the merging process, we use the number of likes on the comments as a weighting factor, applying a weighted calculation to the sentiment scores between subgroups involved in the sentiment. This results in a total sentiment score between the subgroups. Since not all triplets have a personal stance, we complete them by approximating the occurrence frequency of all known personal stances as probabilities and use these probabilities to fill in the incomplete triplets. These operations result in the final CSN.

Initialize:
Initialize 10x10 matrices $adjMatrix$, $weightSum$ and $countMatrix$ to zero
Initialize 1x10 vector $commentCount$ to zero
$incompleteTriplets \leftarrow []$
Process Complete Triplets:
for each $triplet$ in $triplets$ **do**
    if $triplet.group$ is not null **then**
        $src \leftarrow triplet.group$
        $tgt \leftarrow triplet.targetGroup$
        $weightedScore \leftarrow triplet.score \times \max(triplet.likes, 1)$
        $adjMatrix[src][tgt] \leftarrow adjMatrix[src][tgt] + weightedScore$
        $weightSum[src][tgt] \leftarrow weightSum[src][tgt] + \max(triplet.likes, 1)$
        $countMatrix[src][tgt] \leftarrow countMatrix[src][tgt] + 1$
        $commentCount[src] \leftarrow commentCount[src] + 1$
    else
        $incompleteTriplets.add(triplet)$
    end if
end for
Process Incomplete Triplets:
for each $triplet$ in $incompleteTriplets$ **do**
    $tgt \leftarrow triplet.targetGroup$
    $probabilities \leftarrow []$
    for $i \leftarrow 1$ to 10 **do**
        $probabilities[i] \leftarrow countMatrix[i][tgt] / \sum_{j=1}^{10} countMatrix[j][tgt]$
    end for
    $src \leftarrow$ sample a group based on $probabilities$
    $weightedScore \leftarrow triplet.score \times \max(triplet.likes, 1)$
    $adjMatrix[src][tgt] \leftarrow adjMatrix[src][tgt] + weightedScore$
    $weightSum[src][tgt] \leftarrow weightSum[src][tgt] + \max(triplet.likes, 1)$
    $countMatrix[src][tgt] \leftarrow countMatrix[src][tgt] + 1$
    $commentCount[src] \leftarrow commentCount[src] + 1$
end for
Compute Averages:
for $i \leftarrow 1$ to 10 **do**
    for $j \leftarrow 1$ to 10 **do**
        if $weightSum[i][j] > 0$ **then**
            $adjMatrix[i][j] \leftarrow adjMatrix[i][j] / weightSum[i][j]$
        end if
    end for
end for
Output: Draw the group affect network using $adjMatrix$

Figure 4. Construction of community sentiment network

Variable	Explanation
weightSum	triplets' weight
countMatrix[i][j]	number of triplets whose person stance is i and target subgroup is j
incompleteTriplets	variable for storing incomplete triplets
commentCount	number of comments in every subgroup

Table 1. The explanation of some variables in Figure 4

Community opposition index (COI)

We have designed a dedicated group polarization metric for CSN to derive an interpretable polarization index from its complex graph structure. In previous research on sentiment-based polarization measurement, a widely accepted viewpoint is that the stronger the internal cohesion within subgroups and the greater the hostility between subgroups, the higher the level of polarization (Iyengar et al., 2012; Iyengar et al., 2015; Lelkes et al.,2017; Wakefield et al., 2023; Yarchi et al., 2021). The *sentiment thermometer* was developed based on this perspective, and its approach of calculating sentiment temperature differences has gained widespread recognition and practical use (Boxell et al., 2020; Iyengar et al., 2019; Iyengar et al., 2012; Lelkes et al.,2017; Wakefield et al., 2023). However, as we mentioned earlier, this method is only applicable when there are exactly two subgroups involved in group polarization. Therefore, we extended this calculation method to the multi-group domain and proposed the community opposition index (COI).

Firstly, we calculate the sentiment score of a subgroup towards the other subgroups:

$$\left(-e_{ij}\right) \cdot 1_{e_{ij} \leq 0}.$$

Here, e_{ij} represents the sentiment score of subgroup i towards subgroup j. $1_{eij \leqslant 0}$ means that we consider friendly subgroups as not contributing to the overall group polarization.

Subsequently, we sum the sentiment scores of subgroup i towards all other related subgroups and take into account the internal cohesion within each subgroup. Therefore, we get the polarization score of subgroup i. Here, t_i represents the internal cohesion of the subgroup i:

$$t_i \cdot \sum_j \left(-e_{ij}\right) \cdot 1_{e_{ij} \leq 0}.$$

Finally, we weight the overall sentiment score of each subgroup according to its size and calculate the final polarization score:

$$\sum_i \left(\frac{n_i}{N} \cdot t_i \cdot \sum_j \left(-e_{ij}\right) \cdot 1_{e_{ij} \leq 0} \right).$$

Here N represents the total number of comments on the target topic and n_i represents the number of comments of subgroup i within this topic.

It is important to emphasize that, since our metric is a relative indicator, it can avoid interference in the polarization measurement results caused by differences in the number of comments. Additionally, this metric, by focusing on both the internal and external sentiments of subgroups, offers better interpretability.

Zero-shot experiments

Our experiments will focus on the multi-agent system. Since there is no established benchmark in the field of group polarization research, we have chosen to test the system using stance detection

tasks, which share a similar nature. However, unlike stance detection, where subgroups are predefined, our system autonomously identifies and extracts subgroups, making our task distinct from stance detection. We describe the specific setup of our experiments as follows.

Datasets

Based on existing work in the field of stance detection, we will conduct our experiments on the following three datasets (Augenstein et al., 2016; Liang et al., 2022):

SEM16 (Mohammad et al., 2016). This dataset includes six different targets selected from various domains, namely Donald Trump (DT), Hillary Clinton (HC), Feminist Movement (FM), Legalization of Abortion (LA), Atheism (A), and Climate Change is a Real Concern (CC). It includes three types of stances: Favour, Against, and None.

P-Stance (Li et al., 2021). This dataset includes six different targets selected from political domains, namely Donald Trump (Trump), Joe Biden (Biden), Bernie Sanders (Sanders). It includes three types of stances: Favour and Against.

VAST (Allaway et al., 2020). This dataset includes large number of varying targets, and it includes three types of stances: Pro, Con and Neutral.

The statistics of our utilized datasets are shown in Table 2. Since our model's use case is almost zero-shot, we will utilize these three datasets to conduct zero-shot stance detection. We will strictly adhere to the licensing requirements of the respective datasets.

To better evaluate the model's performance, we selected appropriate metrics based on existing literature (Allaway et al., 2021; Lan et al., 2024; Liu et al., 2021). For the SEM16 and P-Stance datasets, we chose F_{avg}, which represents the average of F1 scores for Favour and Against. For the VAST dataset, we opted for Macro-F1 as the evaluation metric.

Dataset	Target	Pro	Con	Neutral
SEM16	DT	148 (20.9%)	299 (42.3%)	260 (36.8%)
	HC	163 (16.6%)	565 (57.4%)	256 (26.0%)
	FM	268 (28.2%)	511 (53.8%)	170 (17.9%)
	LA	167 (17.9%)	544 (58.3%)	222 (23.8%)
	A	124 (16.9%)	464 (63.3%)	145 (19.8%)
	CC	335 (59.4%)	26 (4.6%)	203 (36.0%)
P-Stance	Biden	3217 (44.1%)	4079 (55.9%)	-
	Sanders	3551 (56.1%)	2774 (43.9%)	-
	Trump	3663 (46.1%)	4290 (53.9%)	-
VAST	-	6952 (37.5%)	7297 (39.3%)	4296 (23.2%)

Table 2. Statistics of our utilized datasets

Model adjustment

Since the primary purpose of our designed multi-agent system is to construct the CSN, we need to adjust the model for the experiments. We removed the subgroup exploration expert and eliminated the subgroup exploration process. Instead, we input texts with predefined target groups into the remaining five agents and obtained the output from the polarization assessor. The adjusted model's output only includes the sentiment score and target group, making it suitable for performing stance detection tasks.

Regarding the specific details of the model configuration, we use multiple GPT-3.5 Turbo models provided by OpenAI to serve as agents in the background mining stage and semantic analysis stage,

while GPT-4 is employed as the polarization assessor. This selection was primarily based on a balance between cost and the desired final performance.

Comparison methods

We compare our method with various methods in zero-shot stance detection. This includes adversarial learning method: TOAD, contrastive learning methods: PT-HCL, Bert-based techniques: TGANet and Bert-GCN, LLM-based techniques: GPT-3.5 Turbo, GPT-3.5 Turbo + Chain of thought (COT) and COLA (Allaway et al., 2021; Liang et al., 2022; Allaway et al., 2020; Liu et al., 2021; Zhang et al., 2022; Zhang et al., 2023; Lan et al., 2024).

Zero-shot stance detection results

In Table 3, we present the performance of our method on the zero-shot stance detection task, along with a comparison to other baselines. The results demonstrate that our method exhibits excellent performance in this task, with a performance improvement of 8.4% over the current best result on the VAST dataset. Although our method did not achieve state-of-the-art (SOTA) results across all metrics, its ability to come close to or surpass current SOTA algorithms indicates its significant value when applied to group polarization research.

Model	**SEM16**						**P-Stance**			**VAST**
-	DT	HC	FM	LA	A	CC	Trump	Biden	Sanders	All
TOAD	49.5	51.2	54.1	46.2	46.1	30.9	53	68.4	62.9	41
TGA Net	40.7	49.3	46.6	45.2	52.7	36.6	-	-	-	65.7
BERT-GCN	42.3	50	44.3	44.2	53.6	35.5	-	-	-	68.6
PT-HCL	50.1	54.5	54.6	50.9	56.5	38.9	-	-	-	71.6
GPT-3.5	62.5	68.7	44.7	51.5	9.1	31.1	62.9	80	71.5	62.3
GPT-3.5+COT	63.3	70.9	47.7	53.4	13.3	34	63.9	81.2	73.2	68.9
COLA	68.5	81.7	63.4	71	70.8	65.5	86.6	**84**	79.7	73
Ours	**74.4***	**81.9**	**70.3***	**75.8***	**76.9***	**70.7**	**87.9**	83.2	**86.2***	**81.4***

Table 3. Comparison of our method and baselines in zero-shot stance detection task, all values are percentages. Bold refer to the best performance. '*' denotes our method improves the best baseline at $p < 0.05$ with paired t-test

Limitations and future works

A major limitation of this study is the lack of a widely accepted quantitative definition of group polarization, which hinders the development of standardized benchmarks and affects the rigor of our experiments. Additionally, while our proposed CSN framework, as a temporal dynamic graph, has potential for modelling the evolution of group polarization using graph neural networks (GNNs), the absence of benchmarks limits further exploration. Future work should prioritize establishing standardized definitions and benchmarks to enable more rigorous experimentation and predictive modelling.

Conclusion

In this paper, we discussed the shortcomings of current group polarization measurement approaches and proposed our multi-agent and graph-based measurement scheme. Our solution innovatively introduces a large language model-based multi-agent system into the measurement of group polarization and utilizes the community sentiment network (CSN) to represent the polarization state. Additionally, we provided a metric (community opposition index) for calculating polarization levels using the CSN, allowing the polarization state to be quantified. Finally, we tested our multi-agent system through a zero-shot stance detection task, and the results demonstrated its usability and value.

Acknowledgements

We appreciate the support from the School of Information Management at Wuhan University, which enabled us to complete this paper as students.

About the authors

Zixin Liu is an undergraduate student at the School of Information Management, Wuhan University, PRC, majoring in Information Management and Information Systems. His primary research interests include knowledge graphs and graph learning. Contact him at lzx2562521178@outlook.com

Ji Zhang is an undergraduate student at the School of Information Management, Wuhan University, PRC, majoring in Information Management and Information Systems. He is the co-first author.

Yiran Ding is an undergraduate student at the School of Information Management, Wuhan University, PRC, majoring in Information Management and Information Systems.

References

Allaway, E., & McKeown, K. (2020). Zero-shot stance detection: A dataset and model using generalized topic representations. arXiv preprint arXiv:2010.03640. https://arxiv.org/abs/2010.03640

Allaway, E., Srikanth, M., & McKeown, K. (2021). Adversarial learning for zero-shot stance detection on social media. arXiv preprint arXiv:2105.06603. https://arxiv.org/abs/2105.06603

Augenstein, I., Rocktäschel, T., Vlachos, A., & Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. arXiv preprint arXiv:1606.05464. https://arxiv.org/abs/1606.05464

Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2020). Learning political polarization on social media using neural networks. *IEEE Access*, 8, 47177–47187.https://doi.org/10.1109/ACCESS.2020.2972932

Bilal, M., Gani, A., Marjani, M., & Malik, N. (2019). Predicting elections: Social media data and techniques. In 2019 *International Conference on Engineering and Emerging Technologies* (ICEET) (pp. 1-6). Springer. https://doi.org/10.1109/CEET1.2019.8711854

Bomsdorf, E., & Otto, C. (2007). A new approach to the measurement of polarization for grouped data. AStA *Advances in Statistical Analysis*, 91(2), 181–196. https://doi.org/10.1007/s10182-007-0047-x

Boxell, L., Conway, J., Druckman, J. N., & Gentzkow, M. (2020). Affective polarization did not increase during the coronavirus pandemic. National Bureau of Economic Research. https://www.nber.org/papers/w27133

Bravo, R. B., Del Valle, M. E., & Gavidia, À. R. (2015). A multilayered analysis of polarization and leaderships in the Catalan Parliamentarians' Twitter Network. In 2015 *Fifteenth International Conference on Advances in ICT for Emerging Regions* (ICTer) (pp. 200–206). IEEE. https://doi.org/10.1109/ICTER.2015.7377689

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv*:2005.14165. https://arxiv.org/abs/2005.14165

Conover, M. D., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on Twitter. In *Proceedings of the international AAAI conference on web and social media* (Vol. 5, No. 1, pp. 89-96). https://doi.org/10.1609/icwsm.v5i1.14126

Garcia, D., Abisheva, A., Schweighofer, S., Serdült, U., & Schweitzer, F. (2015). Ideological and temporal components of network polarization in online political participatory media. *Policy & Internet*, 7(1), 46–79. https://doi.org/10.1007/s12132-014-0185-7

Gaurav, M., Srivastava, A., Kumar, A., & Miller, S. (2013). Leveraging candidate popularity on Twitter to predict election outcome. In *Proceedings of the 7th workshop on social network mining and analysis* (pp. 1-8). Association for Computing Machinery. https://doi.org/10.1145/2501025.2501038

Guerra, P., Meira Jr, W., Cardie, C., & Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 215-224). https://doi.org/10.1609/icwsm.v7i1.14421

Hart, P. S., Chinn, S., & Soroka, S. (2020). Politicization and polarization in COVID-19 news coverage. *Science Communication*, 42(5), 679–697. https://doi.org/10.1177/1075547020926489

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, 50(6), 1141–1151. https://doi.org/10.1037/0022-3514.50.6.1141

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22(1), 129–146. https://doi.org/10.1146/annurev-polisci-050518-030121

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, 76(3), 405–431. https://doi.org/10.1093/poq/nfs069

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3), 690–707. https://doi.org/10.1111/ajps.12165

Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2018). Predicting elections from social media: A three-country, three-method comparative study. *Asian Journal of Communication*, 29(3), 252–273. https://doi.org/10.1080/01292986.2018.1455448

Jiang, J., Ren, X., & Ferrara, E. (2023). Retweet-BERT: Political leaning detection using language features and information diffusion on social networks. In *Proceedings of the international AAAI conference on web and social media* (Vol. 17, pp. 459-469). https://doi.org/10.1609/icwsm.v17i1.22160

Jiang, S., & Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1–23. https://doi.org/10.1145/3274357

Knox, R. E., & Inkster, J. A. (1968). Postdecision dissonance at post time. *Journal of Personality and Social Psychology*, 8(4, Pt. 1), 319. https://doi.org/10.1037/h0025557

Kogan, N., & Wallach, M. A. (1967). Risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. *Journal of Experimental Social Psychology*, 3(1), 75–84. https://doi.org/10.1016/0022-1031(67)90025-4

Lan, X., Gao, C., Jin, D., & Li, Y. (2024). Stance detection with collaborative role-infused LLM-based agents. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 18, pp. 891-903). https://doi.org/10.1609/icwsm.v18i1.31360

Lelkes, Y., & Westwood, S. J. (2017). The limits of partisan prejudice. The Journal of Politics, 79(2), 485-501. https://doi.org/10.1086/688223

Li, A., Liang, B., Zhao, J., Zhang, B., Yang, M., & Xu, R. (2023). Stance detection on social media with background knowledge. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 15703-15717). https://doi.org/10.18653/v1/2023.emnlp-main.972

Liang, B., Chen, Z., Gui, L., He, Y., Yang, M., & Xu, R. (2022). Zero-shot stance detection via contrastive learning. In Proceedings of the ACM Web Conference 2022 (pp. 2738-2747). https://doi.org/10.1145/3485447.3511994

Liu, R., Lin, Z., Tan, Y., et al. (2021). Enhancing zero-shot and few-shot stance detection with common sense knowledge graph. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 3152-3157). https://doi.org/10.18653/v1/2021.findings-acl.278

Li, Y., Sosea, T., Sawant, A., Nair, A. J., Inkpen, D., & Caragea, C. (2021). P-stance: A large dataset for stance detection in political domain. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 2355-2365). https://doi.org/10.18653/v1/2021.findings-acl.208

Lu, H. C., & Lee, H. W. (2024). Agents of Discord: Modeling the Impact of Political Bots on Opinion Polarization in Social Networks. Social Science Computer Review. Advance online publication. https://doi.org/10.1177/08944393241270382

Maia, H. P., Ferreira, S. C., & Martins, M. L. (2023). Controversy-seeking fuels rumor-telling activity in polarized opinion networks. Chaos, Solitons & Fractals, 169, 113287. https://doi.org/10.1016/j.chaos.2023.113287

Medaglia, R., & Zhu, D. (2017). Public deliberation on government-managed social media: A study on Weibo users in China. Government Information Quarterly, 34(3), 533-544. https://doi.org/10.1016/j.giq.2017.02.002

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 Task 6: Detecting stance in tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 31-41). https://doi.org/10.18653/v1/S16-1003

Ribeiro, M. H., Calais, P. H., Almeida, V. A., & Meira, W., Jr. (2017). "Everything I disagree with is #FakeNews": Correlating political polarization and spread of misinformation. arXiv preprint arXiv:1706.05924. https://arxiv.org/abs/1706.05924

Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. Proceedings of the National Academy of Sciences, 118(50), e2102141118. https://doi.org/10.1073/pnas.2102141118

Stoner, J. A. F. (1961). A comparison of individual and group decisions involving risk (Doctoral dissertation, Massachusetts Institute of Technology).

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 4, No. 1, pp. 178-185). https://doi.org/10.1609/icwsm.v4i1.14009

Tyagi, A., Uyheng, J., & Carley, K. M. (2020). Affective Polarization in Online Climate Change Discourse on Twitter. In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 443-447). IEEE. https://doi.org/10.1109/ASONAM49781.2020.9381419

Vicario, M. D., Gaito, S., Quattrociocchi, W., Zignani, M., & Zollo, F. (2017). News Consumption during the Italian Referendum: A Cross-Platform Analysis on Facebook and Twitter. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 648-657). IEEE. https://doi.org/10.1109/DSAA.2017.33

Wakefield, R. L., & Wakefield, K. (2023). The antecedents and consequences of intergroup affective polarization on social media. Information Systems Journal, 33(3), 640-668 https://doi.org/10.1111/isj.12319

Xiao, Z., Song, W., Xu, H., Ren, Z., & Sun, Y. (2020). TIMME: Twitter ideology-detection via multi-task multi-relational embedding. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2258-2268). https://doi.org/10.1145/3394486.3403275

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. Political Communication, 38(1-2), 98-139. https://doi.org/10.1080/10584609.2020.1847458

Zhang, B., Ding, D., & Jing, L. (2022). How would stance detection techniques evolve after the launch of ChatGPT? arXiv preprint arXiv:2212.14548. https://arxiv.org/abs/2212.14548

Zhang, B., Fu, X., Ding, D., Huang, H., Li, Y., & Jing, L. (2023). Investigating chain-of-thought with ChatGPT for stance detection on social media. arXiv preprint arXiv:2304.03087. https://arxiv.org/abs/2304.03087

Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 793-803). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330961

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

The role of ontologies in machine learning: a case study of gene ontology

Qiaoyi Liu and Jian Qin
DOI: https://doi.org/10.47989/ir30iConf47575

Abstract

Introduction. Ontologies as knowledgebases have been heavily applied in computational biological studies by implementing into ML models for purposes such as disease-gene associations identification.

Method. We conduct a case study using gene ontology (GO) annotation data and three ML models to replicate the prediction of autism spectrum disorder (ASD)-causing genes.

Analysis. Data were collected from GO and Simmons Foundation Autism Research Initiative (SFARI). The semantic similarities between GO annotation terms on gene products were calculated.

Results. The best-performing model can reach an AUC of .85, which means using GO annotation data for ASD disease-gene prediction can receive a significantly accurate result. However, we stress the importance of constructing knowledgebases in adapting to LLMs and the role of LIS professionals in curating community knowledge for interoperability and reuse.

Conclusion. Biomedical ontologies play a crucial role in the discovery of biomedical knowledge. Knowledge organization and computer science domains require more communication and synchronization in the face of emerging AI and ML technologies.

Introduction

Ontologies as knowledge representation and organization systems have had a prolific increase in the last thirty years. This is particularly the case in the biomedical field. As of this writing, BioPortal, the world's most comprehensive repository of biomedical ontologies, have registered 1,145 ontologies that contain a total of 15.6 million classes, 36,286 properties, and 99.5 million mappings (National Center for Biomedical Ontology, 2024). The vast amount of structured data resulted from these ontologies have been utilized by researchers to explore disease-gene relations and predict genes that have high risk to specific diseases such as Alzheimer's disease (Asif et al., 2018) (Yang et al., 2021) (Binder et al., 2022). The fast growth of biomedical data with parallel advancement of machine learning (ML) algorithms and computational tools created the necessary condition for scientists to respond to urging problems such as identification of genomics for disease research (Piñero et al., 2020) (Krishnan et al., 2016), protein structure analysis (Radivojac et al., 2008), phylogenetic inference (Ata et al., 2021).

The well-structured and actively curated data in knowledge organization systems provide the quality and structures desired by ML model developers, especially in the shift from model-centric artificial intelligence (AI) to data-centric AI (Ng, 2024). However, there has been a lack of communication and understanding between different communities of information science, computer science, related scientific disciplines on the role of knowledge organization systems (KOS), the work involved in building them, and how KOS may be transformed into knowledgebases for AI modelling and applications. The work involved in building KOS includes not only defining subject terms and mapping controlled terms to free-text keywords, but more importantly, defining semantic relations between entities/concepts, bridging raw and machine-readable data, and curating and organizing structured and interoperable data through human intervention and/or automatic methods. This type of knowledge work is essential in bringing scattered, natural language information into systematic representations of knowledge, i.e. data with interpretations (Haendel et al., 2018) for machine to process and even conduct automated reasoning. Despite ontologies and knowledgebases are being constructed across disciplinary fields, the theoretical and methodological aspects of these knowledge organization and representation systems have been largely confined in separate communities, which hinder the communication and sharing of research across communities who study and build KOS.

The purpose of this paper is to elucidate the role of KOS as a trustable data source for AI/ML modelling and applications. We will start with reviewing the current developments of ontologies and knowledgebases, highlighting how knowledgebase structure and design can improve its functions in a ML research workflow. Using a case study of gene ontology (GO), we will illustrate the role that gene ontology played in building ML prediction models. In this case study the GO annotation data and supervised ML models are used to calculate the functional similarities of autism syndrome disease (ASD) genes. The detail of methods and results of the case study is presented in Section 3, followed by Sections 4 and 5, where we speculate on the practices and factors contributing to the quality of GO annotation data. Future KOS research direction is discussed on processing multiple data formats, e.g., images and data sources, disease ontologies, and clinical database, to accommodate AI/ML models.

Evolving knowledge organization systems

There are four main types of KOS. Based on the level of sophistication, the simplest type is term lists, such as glossaries and dictionaries. The second type is metadata-like models, including gazetteers, directories, and authority files. Classification and categorization go to the next level of sophistication because of the embedded relations between concepts or classes in

classification/categorization schemes, taxonomies, and subject headings. The relationship models are the most sophisticated among the four KOS types. Ontologies and semantic networks are the two members of the relationship model, which possess all traits of a KOS that one can hope for: explicitly representing concepts/entities with unambiguous terms while embedding relations between concepts/entities (Zeng, 2008). In the sense of representing a conceptual system via a logical theory, an ontology consists of an annotated and indexed set of formal propositions or assertions about things, a collection of assertions that are called a theory in logic (Guarino & Giaretta, 1995). This special property of ontologies is as closely as it can be in fitting the knowledge representation ideology in AI, which emphasizes sufficiently precise notation, adequacy and expressiveness of representation schemes (Bench-Capon, 1990).

Ontologies sometimes are also called knowledgebases because they are essentially a collection of symbolic structures representing the world based on our cognition (Levesque & Lakemeyer, 2022). Such representation is semantically rich with not only unambiguous vocabularies for entities and individuals in the entities, but also explicit relations between the entities that go beyond hierarchical and associative relations that are the only available relation types in many traditional KOS. Not all ontologies, however, can be called knowledgebases. For example, Schema.org is an ontology developed in collaboration among Google, Microsoft, Yahoo, and Yandex to be used for representing web content as structured data (Google et al., 2024). The purpose of Schema.org is to provide a standard representation scheme for classes of entities and properties these entities possess. Since it is just a representation scheme, it does not contain data (i.e., individual members of classes). Therefore, it is not a knowledgebase and cannot be used for modelling or reasoning purposes.

While many traditional KOS are not designed for problem-solving nor as a data source for ML modelling purposes, many ontologies in the biomedical domain have disrupted the tradition and evolved into knowledgebases. For instance, gene ontology (GO) (Ashburner et al., 2000) and disease ontology (DO) (Bello et al., 2018) use axioms to model knowledge in order to define the semantic interpretation of the presented entities, rules, and class constraints and present multiple relations between concepts. Knowledge organization and AI communities, two research fields that were once separated and did not have much communication before, are now drawing closer and converging through advances in semantic web technologies and data science (Qin, 2020). This trend symbolizes a change to knowledge organization practices. More importantly, it is a signal to knowledge organization moving towards a more interoperable, practical, and heterogeneous identity that exceeds the simple purpose of storing knowledge.

One application of biomedical ontologies is in disease-gene association discovery and identification. Advanced genome sequencing technologies accelerated the process of exploring genomic variations (Piñero et al., 2020) and genetic markers' detection (Chang et al., 2024), generating vast amount of data. This data is preserved and organized into KOS by biocurators who apply knowledge organization theories and practices, which creates semantically rich data sources for applying ML algorithms to analyse larger and complex data sets (Libbrecht & Noble, 2015). Complex diseases with a strong genetic influence often have multiple aetiologies with the involvement of possibly hundreds of different genes. Supervised ML methods can trace hidden relationships among disease-causing genes in existing datasets to discriminate unknown disease genes from non-disease genes (Asif et al., 2018). This advancement plays a crucial role in disease diagnosis and form the basis for clinical decision-making (Chang et al., 2024). This approach was once difficult to proceed due to the lack of: (a) structured KOS with semantically rich relationships between properties; and (b) a powerful computational hardware with feasible models to analyse

large heterogeneous data sets. Problem (a) was greatly improved through making data FAIR (findable, accessible, interoperable, reusable) (Wilkinson et al., 2016) and encoding languages (e.g. XML and JSON) for representing machine-readable knowledge and reflecting reality. Problem (b) now is being addressed by numerous supervised and unsupervised ML models that are developed and applied to assist in areas such as genetics and molecular science.

Case study: using GO annotation and ML models to identify autism spectrum disorder (ASD)

Case selection

Gene ontology (GO)

The fast-developing nucleotide sequencing techniques and gene expression analysis has urged the biological community to establish a knowledge resource for this massive data. Unlike other STEM research fields, biological knowledge can be less explained by mathematical equations but more through natural language (Pesquita et al., 2009). Gene ontology (GO, http://geneontology.org) constructed by the gene ontology consortium, is crowned as the GOld mine. It provides "a comprehensive, structured, computer-accessible representation of gene function for genes from any cellular organism or virus." Until 2023, GO contains 43,303 biological terms as annotations to gene products, linked together by 88,099 relationships (The Gene Consortium, 2023). The GO knowledgebase represents a standardized controlled vocabulary which defines various components of molecular biology shared amongst life forms (Yousef et al., 2021). It has become a critical component of life science research, supporting analysis of large-scale genomics data analysis and biological systems (Duck et al., 2016) and broadly used in research, clinical diagnosis, and industry. Over the years many ML models and equations were developed specifically for processing and using GO data. Considering its significance to biological research and established computational techniques, we select GO annotation data as our primary data source to conduct case study.

GO construction and GO annotations

Bio-entities described in GO can be considered as knowledge unites, or 'things' such as gene products. The entities in GO are structured as a directed acyclic graph (DAG) in which GO terms/annotations are represented as nodes and relationships between terms are represented as edges that follow certain directions and never form a closed loop (Asif et al., 2018). Each ontology term (called 'class' in the field of ontology) represents a functional characteristic that can be attributed to a gene product (The Gene Consortium, 2023). Terms representing the 'things' are organized into three categories: molecular function (MF), biological process (BP), and cellular component (CC). Each term is described by five required elements: a unique ID, term name, aspect (which category it belongs to), definition, and relationships to other terms. GO links the terms by using a set of triple statements, most commonly '*is_a*' or '*part_of*', which stands for class-subclass relationship and part-whole relationship, respectively (see table 1). A GO annotation is an association between a specific gene product and a GO term and should be interpreted as a statement that the gene product possesses the functional characteristics represented by the GO term. Each GO annotation covers only one characteristic of the gene product. Therefore, a gene product can have multiple GO annotations. GO annotations are continually added to the knowledgebases from 173,000 scientific papers. All annotations are supported by an evidence code which describes the type of evidence and a reference that lists a persistent identifier for tracing the source. For quality control, they are regularly reviewed, edited, or removed by biocurators or the GO user community (The Gene Consortium, 2023).

Relation	Description	Example
is_a	The basic structure of GO. If we say A *is_a* B, we mean that entity A is a subtype of entity B.	Mitotic cell cycle *is_a* cell cycle, or lyase activity *is_a* catalytic activity.
part_of	The *part of* relation is used to represent part-whole relation. A *part of* relation would only be added between A and B if B is **necessarily** *part of* A: wherever B exists, it is as *part of* A, and the presence of the B implies the presence of A.	If a gene product X is annotated as located in the inner mitochondrial membrane and the ontology records a *part of* relation between inner mitochondrial membrane and mitochondrion, we can safely conclude that X is located in a mitochondrion.
has_part	The logical complement to the *part of* relation is *has part*, which represents a part-whole relationship from the perspective of the parent. As with *part of*, the GO relation *has part* is only used in cases where A always has B as a part, i.e., where A necessarily *has part* B. If A exists, B will always exist; however, if B exists, we cannot say for certain that A exists. i.e., all A have part B; some B part of A.	A receptor tyrosine kinase activity *has part* ATP hydrolysis activity. However, it would not then be correct to group all annotations to kinase activity under ATPase activity.
regulates	A relation that describes case in which one process directly affects the manifestation of another process or quality, i.e., the former *regulates* the latter. The target of the regulation may be another process, for e.g., regulation of a pathway or an enzymatic reaction, or it may be a quality, such as cell size or pH. Analogously to *part of*, this relation is used specifically to mean necessarily *regulates*: if both A and B are present, B always *regulates* A, but A may not always be regulated by B., i.e., all B *regulate* A; some A are *regulated by* B.	If gene product X is annotated as involved in a process that *regulates* glycolysis, it would not be correct to conclude that X participates in glycolysis.

Table 1. Main term relations used in GO

Identify disease-gene associations in ASD

Disease-gene association prediction has been in the spotlight of bioinformatics research for over a decade. Scholars and doctors are eager to identify gene mutations, which are the primary causes of genetic diseases. Many large-scale genetic studies provided candidate genes that may cause diseases. However, traditional disease-gene association costs heavy human labor and are highly difficult to process due to genetic heterogeneity. The disease-causing genes are identified by statistical geneticists, where linkage analysis and association studies were conducted on candidate genes based on their likelihood of being involved in a specific disease i.e. gene prioritization algorithms (Radivojac et al., 2008). With more heterogeneous data, it is no longer possible to conduct this type of analysis solely by human labor.

Nowadays, computational approaches can speed up this process and are capable of handling more complex diseases, such as autism spectrum disorder (ASD). ASD is a type of neurodevelopmental syndrome that affects about one in 100 people worldwide (Geschwind & State, 2015). Strong evidence indicates the causes include both genetic and environmental factors (Kim & Leventhal, 2015). Diseases like ASD tend to present a highly heterogeneous genotype, which makes it difficult for biological marker identification. Although ML methods can be used to identify these markers, their performance highly depends on the size and quality of available data (Asif et al., 2018). Therefore, using GO annotation data can provide consistent and clean data for ML models to learn from, thus improving their general performance. Our purpose is not developing a new model or equation, but to replicate this process and discuss the data quality and ontology construction from

the LIS scope outside of these two communities. We hope to highlight the significance of structured knowledgebases for computational biological research and applicability in AI LLMs.

Functional similarity vs. semantic similarity

One major issue in identifying disease-gene association is how to define an unknown gene can cause similar outcomes as a disease-associated gene already known. Biologists found that functional similar genes tend to contribute to similar phenotypes. For instance, etiologically relevant genes disrupted by genetic variants in ASD patients tend to aggregate in specific biological processes (Voineagu & Eapen, 2013). This means disease-causing genes and disease-candidate genes may belong to the same tree path in GO DAG structure (Asif et al., 2018). Gene products that share highly overlapping GO terms may have higher functional similarities. If a gene product is on the same tree branch with gene products that are associated with disease-related GO annotation function terms, it indicates this gene may have a higher probability to be a disease-causing gene. Thus, the comparison of gene products' function similarities is transformed into the comparison of semantic similarities of GO terms. Adopting ontology annotation data can provide a means to compare entities on aspects that otherwise not be comparable.

Typically there are two ways to compare terms in graph-structured ontologies such as GO: edge-based or node-based (Pesquita et al., 2009). Edge-based approaches are based on counting the number of edges in the graph path between two terms (Rada et al., 1989). This can be problematic for biological data because the approach is based on two conditions: (i) nodes and edges in the biological ontology are uniformly distributed, and (ii) edges at the same level in the ontology correspond to the same semantic distance between terms. However, biological knowledge can rarely meet these two conditions where terms at the same tree structure level share the same scale and weight.

Node-based approaches are more commonly accepted in biological domain. There have been several statistical measurements and equations developed. Resnik proposed the Information Content (IC) to quantify the informativeness of a concept c as negative the log likelihood (Resnik, 1999):

$$-\log p(c) \tag{1}$$

However, a drawback of Resnik's method is that it ignores the information contained in the structure of the ontology by only concentrating on the information content of a term derived from the corpus statistics. For biological ontologies, the specificity of a GO term is usually determined by its location in the GO graph. A GO term's semantics (biological meanings) are inherited from all its ancestor terms (Wang et al., 2007). For instance, shown in figure 1, GO:0002839 positive regulation of immune response to tumour cell is a child term of GO:0002418 immune response to tumour cell, the latter is a child term of GO:0002347 response to tumour cell and GO:0006955 immune response, both of which are the child terms of GO:0008150 biological processes. Because GO term is the aggregation of all its parent terms, GO:0002839 positive regulation of immune response to tumour cell should possess the characteristics of both GO:0002418 response to tumour cell and GO:0006955 immune response. Not to mention biomedical ontologies usually have various edge length i.e. edges at the same level convey different semantic distances, various depth i.e. terms at the same level have different level of details, and various node density i.e. some areas of the ontology have a greater density of terms than others (Pesquita et al., 2009).

Wang proposed a metrics that is specifically for encoding biological terms' semantics by aggregating the semantic contributions of all its ancestor terms including itself in the GO graph. Wang believes Resnik's IC approach focus is more applicable for knowledge in natural language

such as bird and crane, forest and graveyard but is not the ideal option for biological knowledge in ontologies (Wang et al., 2007). IC neglects the logic that if two GO terms share the same parent are near the root of the ontology i.e., terms that are more general, they should have larger semantic difference than two terms having the same parent and being far away from the root of the ontology because the latter are more specific terms. GO is constructed in such way that if the child GO term describes the gene product, then all its parent terms must also apply to that gene product (Wang et al., 2007). DO contains logical definitions (axioms) to describe relevant disease drivers, constructed with specific relational ontology (RO) terms, to create a restriction between a DO term and another open biological and biomedical ontology (OBO) Foundry ontology term. Thus, DO has the ability to 'infer' the child terms originated from a parent term based on a known parent-child relationship (Qin & Liu, 2024). Therefore, it is reasonable to aggregate the biological meanings of all its ancestor terms when determining its semantic similarity with another GO term.

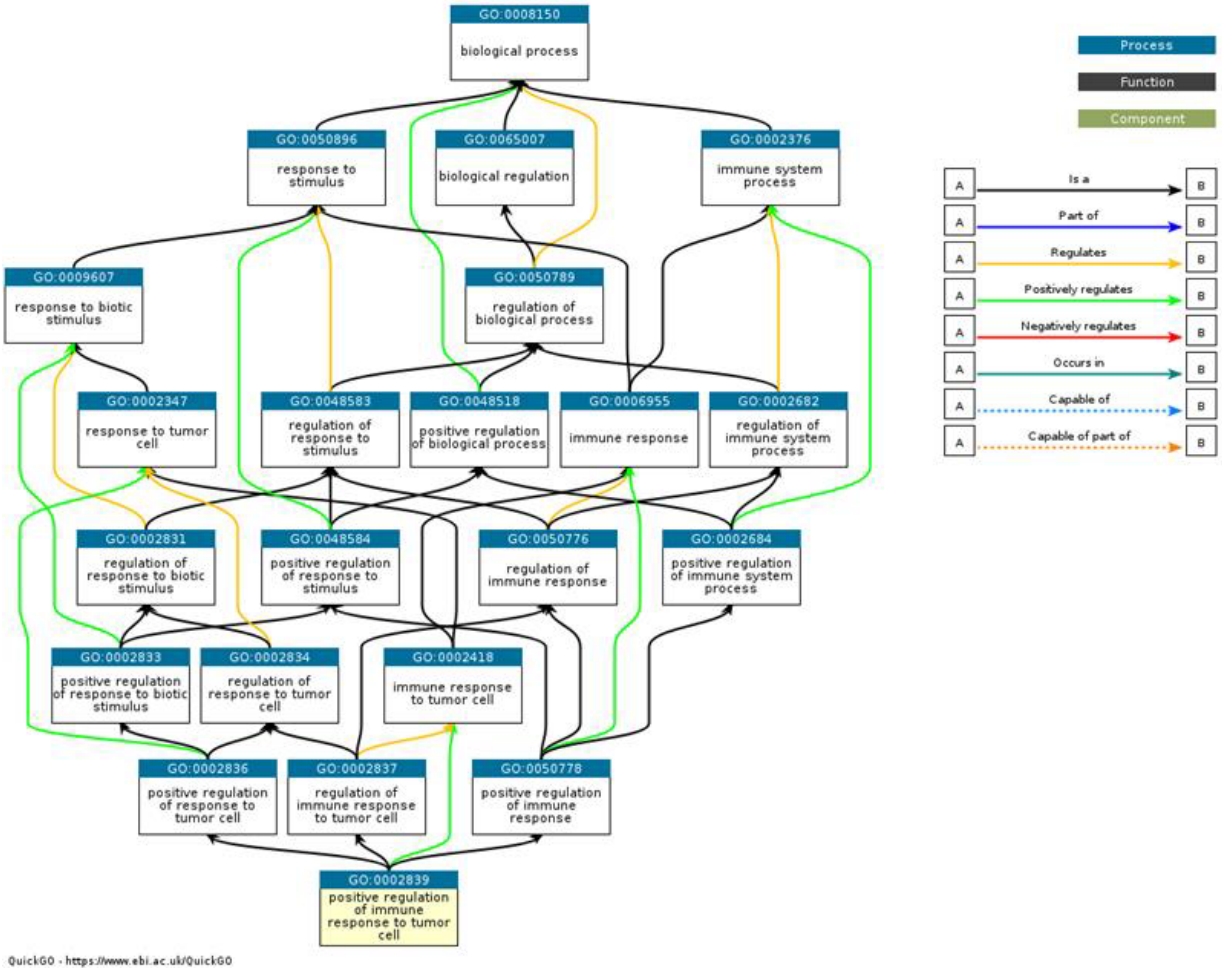


Figure 1. A graphic view of GO:0002939 positive regulation of immune response to tumour cell and all its ancestor terms in GO

Next, the issue is transforming semantic similarity of GO terms into the functional similarity of gene products. Since one gene product may have more than one GO annotation, we must consider the contributions from the semantically similar terms that annotate the genes separately (Wang et al., 2007). Wang first defines the maximum semantic similarity between one GO term and a set of GO = $\{go_1, go_2, \dots, go_k\}$as:

$$Sim(go, GO) = \max_{1 \le i \le k} (S_{GO}(go, go_i)) \quad (1)$$

Then, given two gene products G1 and G2 annotated by sets of GO terms: GO1 = $\{go_{11}, go_{12}, \dots, go_{1m}\}$ and GO2 = $\{go_{21}, go_{22}, \dots, go_{2n}\}$ respectively, the functional similarity of G1 and G2 can be defined as:

$$Sim(G_1, G_2) = \frac{\sum_{1 \le j \le m} Sim(go_{1j}, GO_2) + \sum_{1 \le j \le n} Sim(go_{2j}, GO_1)}{m+n} \quad (2)$$

In this study, we applied Wang's approach to calculate the functional similarity of gene products for predicting candidate genes that lead to ASD disease. GO annotation data were collected using the *org.Hs.eg.db* R package (Carlson et al., 2019). Only terms for biological processes are selected (N = 28,140). Next, we mapped all the GO terms and their associated gene products (N = 157,247). Since our aim is investigating the construction and quality of GO as a knowledge ontology in biomedical research, we only try to replicate part of the workflow by Asif et al., using three of the ML models – support-vector machines (SVM), random forest (RF), and gradient boosting (GB). We randomly selected 1,000 gene products from the mapping as our train set for the models. For our test set, we also randomly selected 20 candidate gene products obtained from the Simons foundation autism research initiative (SFARI, https://gene.sfari.org/) gene database (N = 1,176), in which 15 are categorized by SFARI as high confidence disease genes (HD) and 5 are categorized as low confidence (LD) disease genes. This categorization is used to compare with results from the ML classification models. Table 2 shows the 20 test gene products and their categories. Before we allow the model to conduct semantics similarity calculation, we removed SFARI's categorization for the models to predict and then verify their overall performance by comparing its results with SFARI's classification. The semantic similarity measures between GO terms were implemented using the *GOSemSim* R package (Yu et al., 2010).

Results

As is shown in Fig. 2, we used the train set gene product and their GO annotation data to run the 1,000×1000 similarity matrix. Next, we conducted the 1000×20 similarity matrix between train set and test set (see figure 3). Based on their functional similarity, we identify which of the 20 test set gene products may be a disease-associated gene. Out of the three models we used, SVC-based classifier trained and test on Wang's semantic similarity matrix outperformed the other classifiers, with AUC value equals 0.85 (see table 3). The difference between RF and GB AUC values were minor, indicating the independence of the methodology to the semantic measure. A recall of 0.25 is considerably low. However, given the highly imbalanced dataset (most genes in train set are non-ASD-associated with only a very small number of genes are ASD-associated) and small number of train dataset used, we believe this is an expected and sufficient result to prove that GO annotation data could be used in gene-disease association identification ML models. Furthermore, the overall performance of MLs is crucially dependent on the quality of GO annotation data and structural relationship between GO terms. Here we only demonstrated one way to calculate the semantic similarity between GO terms. The results can vary depending on which type of approach one applies.

Gene product	Candidate gene category by SFARI (ASD causing is 1, non-ASD causing is 0)
SNTG2	1
BIRC6	1
CNTN4	1
ADORA2A	1
LZTR1	1
WWOX	1
HYDIN	1
RBFOX1	1
TRPM1	1
FAN1	1
CMPK1	1
STIL	1
TAL1	1
BCL9	1
OR4A47	1
CARD16	0
CASP1	0
PCDH17	0
P2RX6	0
AHR	0

Table 2. Selected test set gene products and ASD categories (N=20)

Classifiers	AUC	Recall	F1 Score
SVC	0.85	0.25	0.10
RF	0.43	0.25	0.10
GB	0.50	0.25	0.10

Table 3. The performance of classifiers trained and tested over Wang semantic similarities matric. The Area Under the Curve (AUC) evaluation metric was used to estimate and compare the performance of the classifiers

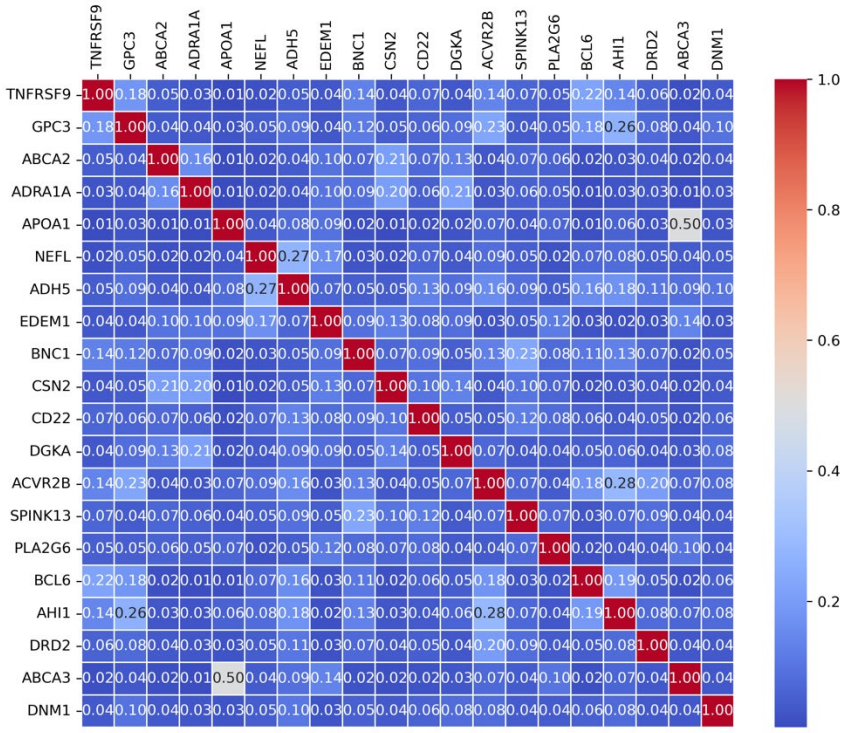


Figure 2. Similarity matrix of the first 20 gene products from train set

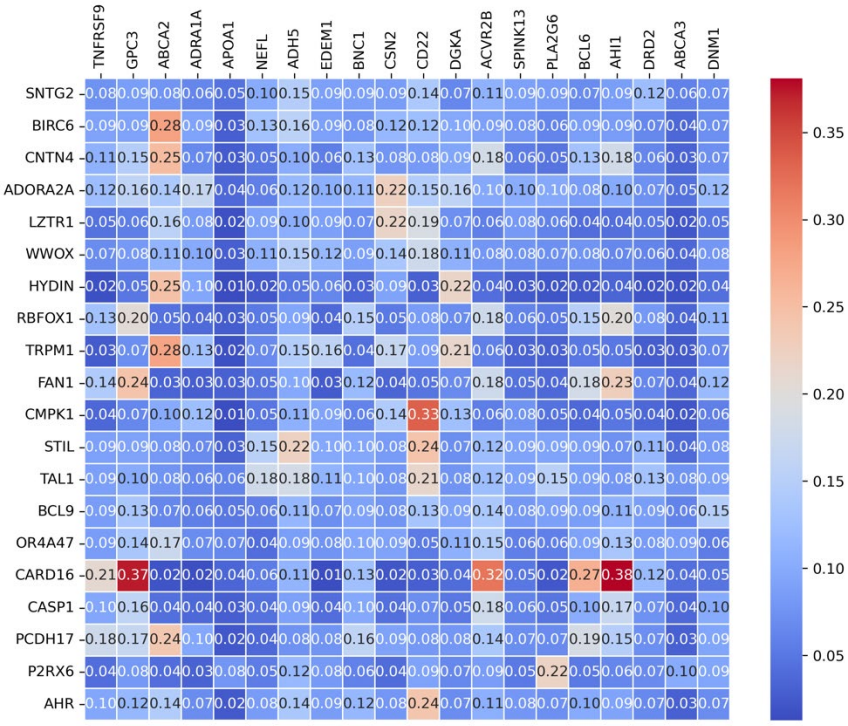


Figure 3. Similarity matrix of the first 20 gene products from train set and all 20 gene products from test set

One major issue which reduces its performance is the imbalance data between one class from other classes. For example, in clinical or disease-related cases, there is inevitably less data from treatment groups than from the normal (control) group (Min et al., 2016). One of the common solutions to this problem is data pre-processing. This is largely conducted by data curators guided by schemas and controlled vocabularies. However, the high proportion of duplicate or near-duplicate samples in biological sequencing data is a serious problem during ML model training which tends to be overlooked (Auslander et al., 2021). Careful data processing is needed to ensure the independence of data between train and test set. When using GO annotation data to quantify the similarity between biological terms, an important issue is that some annotation data in GO is referred from other sources based on similarity. Using these annotations for semantic similarity calculation is in fact, data circularity (Pesquita et al., 2009).

Discussion

Conventional LIS approaches to knowledge representation such as hierarchical and faceted classification integrated human knowledge into a systematic arrangement in which concepts and structures tend to be abstract and have an epistemology orientation. Ontologies, on the contrary, disintegrate parts of knowledge into a problem-solving focused structure and are more pragmatic and application oriented. The disintegrative approach places the entity itself, in our case it is the gene products, in a less important position, but rather, focuses on all aspects related to it i.e., GO terms and annotations (Qin, 2002). As a result, ontologies are more suitable as conceptual frameworks to specific problems. In this case study scenario, GO annotation can be applied to various statistical measurements of semantic similarity, which represents the entity i.e., gene product it describes. The application of GO knowledgebase has transcended beyond its initial purposes which is representing and organizing a phenomenon of a knowledge domain.

As Deep Learning and LLM models are more frequently applied to bioinformatics and biomedical research, we must reconsider schema and frameworks for building knowledge organization systems in order for them to apply to more complex computational approach. Already there have been ways to use LLMs in knowledge graph engineering and ontology construction in replacement of human-conducted natural language processing (Kommineni et al., 2024). Likely we will be witnessing an evolutionary change in knowledge organization and representation under the fast-developing AI era. While computer scientists focus on the pragmatic side of KO and KR by exploiting LLMs and prompt engineering to improve the accuracy, scalability, and depth of knowledge captured, the values in theoretical advancements should not be overlooked. The lack of communication between AI and KOS communities may hinder the applicability of ontologies as knowledge resources. With the domain shifting towards generative AI, more work is necessary to refine the 'core' and paradigm of this interdisciplinary field. This may seem trivial to application-driven science, but a '*step back*' from the actual phenomena is to find a broader characterization that encompasses the instances at hand (Lyytinen et al., 2004). The paradigmatic similarities in KR between KO and AI offer not only theory foundations but also practicalities for KO to contribute its unique value for knowledge representation (Qin, 2020).

Fundamentally, KOSs must be prepared to handle various types of knowledge generated by AI models. Measurements for semantic similarity should be updated to comply with new relations between entities and relationships as new terminologies, or even new knowledge domains, emerge. Guidelines for misinformation detection are necessary if ontologies are using LLM-generated knowledge. Questions on the trustworthiness of AI can also impact the reliability of ontologies (Kaur et al., 2023), especially when the construction of LLMs and algorithms are mostly hidden in 'black boxes'. In terms of medical and health data, the fairness of AI using knowledgebase data can cause ethical issues. Patient privacy and confidentiality are necessary factors that decide whether we are entrusted to use this data in ML models and not be shared for all other purposes.

Applications like disease-gene association identification is closely intertwined with clinical decision-making, which can have a direct impact on medical practice and communication between physicians and patients (Lötsch et al., 2022). A new ethical framework may be in order to balance the need of society and future patients with legitimate expectations of privacy (Haendel et al., 2018), especially with the involvement of AI models.

Conclusion

This paper focuses on using GO as case for training ML models to serve a type of computational biological research – disease-gene association identification. We calculated the functional similarity of gene products represented by the GO annotation semantic similarity, and trained three supervised ML models – SVM, RF, and GB. From experimenting on this workflow, we evaluate the role of ontologies as knowledgebases for large data biomedical research. Applying theories in knowledge organization and representation, we argue that knowledge organization and computer science domains require more communication and synchronization in the face of emerging AI and LLM technologies in order to accommodate to AI-generated knowledge and policies. We conclude that ontologies have played a crucial role in the discovery of biomedical knowledge and clinical decision-making by providing meaningful, structured, and reliable data.

Acknowledgements

We thank the comments provided by the anonymous reviewers for this paper.

About the authors

Qiaoyi Liu is a Ph.D. student in Information Science and Technology at Syracuse University. Her research interests are in knowledge organization (KO) and science of science (SoS). Especially, she studies biological knowledge representation and construction of ontologies guided by classification theories and semantic measurements. She is interested in knowledgebases exploited by ML models and trustworthy LLMs to generate knowledge for bioinformatics and computational biology research. She can be contacted at qliu11@syr.edu

Jian Qin is Professor of the iSchool at Syracuse University. She conducts research in metadata, knowledge modelling and representation, ontologies, research collaboration networks, research impact assessment, and data curation. Jian Qin directs a Metadata Lab, a research group focusing on big metadata analytics and knowledge modelling. Her research has received funding from US NSF, NIH, IMLS, among others. She publishes widely with more than 100 journal and conference papers in the field of information science, scientometrics, knowledge organization, and metadata and been invited to give keynotes, lectures, and presentations at conferences and institutions inside and outside of the U.S. She is the co-author of the book Metadata and co-editor for several special journal issues on knowledge discovery in databases and knowledge representation. She received the 2020 Frederick G. Kilgour Award for Research in Library and Information Technology. Jian Qin holds a Ph.D. from University of Illinois at Urbana-Champaign. She can be contacted at jqin@syr.edu

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. Nature Genetics, 25(1), 25–29. https://doi.org/10.1038/75556

Asif, M., Martiniano, H. F. M. C. M., Vicente, A. M., & Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PLOS ONE, 13(12), e0208626. https://doi.org/10.1371/journal.pone.0208626

Ata, S. K., Wu, M., Fang, Y., Ou-Yang, L., Kwoh, C. K., & Li, X.-L. (2021). Recent advances in network-based methods for disease gene prediction. Briefings in Bioinformatics, 22(4), bbaa303. https://doi.org/10.1093/bib/bbaa303

Auslander, N., Gussow, A. B., & Koonin, E. V. (2021). Incorporating Machine Learning into Established Bioinformatics Frameworks. International Journal of Molecular Sciences, 22(6), Article 6. https://doi.org/10.3390/ijms22062903

Bello, S. M., Shimoyama, M., Mitraka, E., Laulederkind, S. J. F., Smith, C. L., Eppig, J. T., & Schriml, L. M. (2018). Disease Ontology: Improving and unifying disease annotations across species. Disease Models & Mechanisms, dmm.032839. https://doi.org/10.1242/dmm.032839

Bench-Capon, T. J. M. (1990). Knowledge representation: An approach to artificial intelligence. Academic Press.

Binder, J., Ursu, O., Bologa, C., Jiang, S., Maphis, N., Dadras, S., Chisholm, D., Weick, J., Myers, O., Kumar, P., Yang, J. J., Bhaskar, K., & Oprea, T. I. (2022). Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Communications Biology, 5(1), 125. https://doi.org/10.1038/s42003-022-03068-7

Carlson, M., Falcon, S., Pages, H., & Li, N. (2019). Org. Hs. Eg. Db: Genome wide annotation for Human. R Package Version, 3(2), 3.

Chang, J., Wang, S., Ling, C., Qin, Z., & Zhao, L. (2024). Gene-associated Disease Discovery Powered by Large Language Models (No. arXiv:2401.09490). arXiv. https://doi.org/10.48550/arXiv.2401.09490

Duck, G., Nenadic, G., Filannino, M., Brass, A., Robertson, D. L., & Stevens, R. (2016). A Survey of Bioinformatics Database and Software Usage through Mining the Literature. PLOS ONE, 11(6), e0157989. https://doi.org/10.1371/journal.pone.0157989

Geschwind, D. H., & State, M. W. (2015). Gene hunting in autism spectrum disorder: On the path to precision medicine. The Lancet Neurology, 14(11), 1109–1120. https://doi.org/10.1016/S1474-4422(15)00044-7

Google, Microsoft, Yahoo, & Yandex. (2024). Schema.org [Organization]. Schema.Org. https://schema.org/

Guarino, N., & Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. In Towards Very Large Knowledge Bases. (pp. 25–32). IOS Press. https://www.loa.istc.cnr.it/old/Papers/KBKS95.pdf

Haendel, M. A., Chute, C. G., & Robinson, P. N. (2018). Classification, Ontology, and Precision Medicine. The New England Journal of Medicine, 379(15), 1452–1462. https://doi.org/10.1056/NEJMra1615014

Kim, Y. S., & Leventhal, B. L. (2015). Genetic Epidemiology and Insights into Interactive Genetic and Environmental Effects in Autism Spectrum Disorders. Biological Psychiatry, 77(1), 66–74. https://doi.org/10.1056/NEJMra1615014

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., & Troyanskaya, O. G. (2016). Genome-wide prediction and functional characterization of

the genetic basis of autism spectrum disorder. Nature Neuroscience, 19(11), 1454–1462. https://doi.org/10.1038/nn.4353

Levesque, H. J., & Lakemeyer, G. (2022). The logic of knowledge bases (Second edition). College Publications.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332. https://doi.org/10.1038/nrg3920

Min, S., Lee, B., & Yoon, S. (2016). Deep learning in bioinformatics. Briefings in Bioinformatics, bbw068. https://doi.org/10.1093/bib/bbw068

National Center for Biomedical Ontology. (2024). Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies. [Organization]. BioPortal. https://bioportal.bioontology.org/

Ng, A. (Director). (2024, July 4). A Chat with Andrew on MLOps: From Model-Centric to Data-Centric AI[OL]. [Video recording]. https://www.youtube.com/watch?v=06-AZXmwHjo

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. PLOS Computational Biology, 5(7), e1000443. https://doi.org/10.1371/journal.pcbi.1000443

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Research, 48(D1), D845–D855. https://doi.org/10.1093/nar/gkz1021

Qin, J. (2002). Evolving Paradigms of Knowledge Representation and Organization: A Comparative Study of Classification, XML/DTD, and Ontology. ADVANCES IN KNOWLEDGE ORGANIZATION, 8, 465–471.

Qin, J. (2020). Knowledge Organization and Representation under the AI Lens. Journal of Data and Information Science, 5(1), 3–17. https://doi.org/10.2478/jdis-2020-0002

Qin, J., & Liu, Q. (2024). Organizing Knowledge in Knowledgebases: A Case Study. Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities, 393–400.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1), 17–30. https://doi.org/10.1109/21.24528

Radivojac, P., Peng, K., Clark, W. T., Peters, B. J., Mohan, A., Boyle, S. M., & Mooney, S. D. (2008). An integrated approach to inferring gene–disease associations in humans. Proteins: Structure, Function, and Bioinformatics, 72(3), 1030–1037. https://doi.org/10.1002/prot.21989

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 11, 95–130. https://doi.org/10.1613/jair.514

The Gene Consortium. (2023). The Gene Ontology knowledgebase in 2023. Gene, 224(1), 1–14. https://doi.org/10.1093/genetics/iyad031

Voineagu, I., & Eapen, V. (2013). Converging Pathways in Autism Spectrum Disorders: Interplay between Synaptic Dysfunction and Immune Responses. Frontiers in Human Neuroscience, 7. https://doi.org/10.3389/fnhum.2013.00738

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics, 23(10), 1274–1281. https://doi.org/10.1093/bioinformatics/btm087

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Yang, K., Lu, K., Wu, Y., Yu, J., Liu, B., Zhao, Y., Chen, J., & Zhou, X. (2021). A network-based machine-learning framework to identify both functional modules and disease genes. Human Genetics, 140(6), 897–913. https://doi.org/10.1007/s00439-020-02253-0

Yousef, M., Sayıcı, A., & Bakir-Gungor, B. (2021). Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis. In G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, & S. Khan (Eds.), Database and Expert Systems Applications—DEXA 2021 Workshops (pp. 205–214). Springer International Publishing. https://doi.org/10.1007/978-3-030-87101-7_20

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. Bioinformatics, 26(7), 976–978. https://doi.org/10.1093/bioinformatics/btq064

Zeng, M. L. (2008). Knowledge Organization Systems (KOS). KNOWLEDGE ORGANIZATION, 35(2–3), 160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Finding Pareto trade-offs in fair and accurate detection of toxic speech

Soumyajit Gupta, Venelin Kovatchev, Anubrata Das, Maria De-Arteaga, and Matthew Lease
DOI: https://doi.org/10.47989/ir30iConf47572

Abstract

Introduction. Optimizing NLP models for fairness poses many challenges. Lack of differentiable fairness measures prevents gradient-based loss training or requires surrogate losses that diverge from the true metric of interest. In addition, competing objectives (e.g., accuracy vs. fairness) often require making trade-offs based on stakeholder preferences, but stakeholders may not know their preferences before seeing system performance under different trade-off settings.

Method. We formulate the GAP loss, a differentiable version of a fairness measure, Accuracy Parity, to provide balanced accuracy across binary demographic groups.

Analysis. We show how model-agnostic, *HyperNetwork* optimization can efficiently train arbitrary NLP model architectures to learn *Pareto-optimal* trade-offs between competing metrics like predictive performance vs. group fairness.

Results. Focusing on the task of toxic language detection, we show the generality and efficacy of our proposed GAP loss function across two datasets, three neural architectures, and three fairness loss functions.

Conclusion. Our GAP loss for the task of TL detection demonstrates promising results - improved fairness and computational efficiency. Our work can be extended to other tasks, datasets, and neural models in any practical situation where ensuring equal accuracy across different demographic groups is a desired objective.

Introduction

Toxic language in social media is often associated with various risks and harms: cyber bullying, discrimination, mental health, and even hate crimes. Given the massive volume of user-generated content online, manual review of all posts by human moderators simply does not scale. Consequently, natural language processing (NLP) methods have been developed to fully or partially automate toxicity detection (Schmidt & Wiegand, 2017). Prior work has achieved high Accuracy and F1 scores on toxicity detection (e.g., (Zampieri et al., 2020)) across various model architectures:

e.g., convolutional (CNN) (Gambäck & Sikdar, 2017), sequential (BiLSTM) (Graves et al., 2005), and transformer (BERT) (Devlin et al., 2018). However, studies have also found that model accuracy can vary greatly across sensitive demographic attributes, such as race or gender (Das et al., 2021; Park et al., 2018; Sap et al., 2019). Subjective annotation in such tasks arise from personal biases and experiences of annotators. Traditional approaches relying on majority voting to resolve disagreements leads to oversimplification of the task. For example, a BERT-based classifier obtains 90.4% vs. 84.5% accuracy for White vs. African American author on Davidson's dataset (Davidson et al., 2017) when just optimized for overall accuracy, independent of author groups. Thus, in subjective domains, the minority viewpoint plays an important role (Sang & Stanton, 2022), where context and interpretations around data collections (Rahman et al., 2022), sources (Chaudhry & Lease, 2022) and targets, can heavily influence judgments.

While recent years have seen rapid progress in fairness research, it is often measured in a post hoc manner, and optimization is often indirect (*e.g.*, by improving training data through pre- or post-processing) (Sap et al., 2019). A particular challenge is that most existing measures are non-differentiable and thus cannot be optimized directly via gradient descent. While one can optimize differentiable surrogate loss functions instead, this risks *metric divergence* between the optimization criteria used in training *vs.* the actual metrics of interest (Metzler & Croft, 2007; Morgan et al., 2004; Swezey et al., 2021; Yue et al., 2007).

As (Friedler et al., 2021) and others have noted, different worldviews lead to conflicting definitions of fairness that are mutually incompatible and specific fairness measures must be selected (suitable to the given task, context, and stakeholders at hand). In this work, we adopt a popular fairness objective, Accuracy Parity (Zhao et al., 2020) to optimize a model to provide balanced accuracy across demographic groups (Berk et al., 2021; Das et al., 2021; Heidari et al., 2019; Mitchell et al., 2021). Because no differentiable version of this measure exists, we formulate a novel, differentiable version, *Group Accuracy Parity* (GAP) that can be directly used to optimize descent-based models. We provide both a theoretical derivation and an empirical justification for GAP.

However, optimizing GAP alone may reduce Overall Accuracy (OA) since seeking to better fit minority group may lead to worse fit of majority group that tends to drive OA. Ultimately, we face a trade-off between competing objectives, whether we balance between competing accuracy goals (*e.g.*, precision *vs.* recall), fairness goals, or any combination thereof. Multi-Objective Optimization (MOO) provides a principled framework and rigorous toolbox for approaching such competing trade-offs, instead of treating them as single objective regularization problems (Little, 2023; Sorensen et al., 2024; Soto et al., 2022; Suau et al., 2024). We believe such MOO work remains underexplored in NLP today, and to the best of our knowledge, ours is the first NLP work on MOO for fair toxic language detection.

Because competing objectives typically lack global optima, optimization requires choosing among a set of equally valid, *Pareto-optimal* trade-offs between objectives. Naturally, selection of a suitable trade-off depends on stakeholder needs, and they typically wish to see system performance under real trade-off conditions before having to commit to any particular trade-off. We demonstrate how the full Pareto manifold – for *any* underlying model architecture – can be efficiently induced, provided optimization can be performed via gradient descent (with

differentiable loss objectives). This is accomplished via recent advances in *Pareto front learning* (PFL) (Gupta et al., 2022; X. Lin et al., 2020; Navon et al., 2021) for *HyperNetworks* (Ha et al., 2017), which train one neural model to generate effective weights for a second, target model.

In summary, we pursue two distinct and complementary approaches for fair toxic language detection via model optimization. First, recognizing the repeated call for balancing accuracy across demographic groups, yet finding no differentiable metric doing so, we present the first differentiable version, GAP, enabling optimization for the first time via standard gradient descent. Our results show a clear benefit of optimizing directly for the target metric of interest rather than surrogate loss functions that diverge from it. Second, to demonstrate generality of PFL optimization over competing objectives, we induce the full Pareto front of optimal trade-offs between OA *vs.* three different fairness measures: GAP and two prior measures. To show generality of both techniques – single-objective GAP and multiobjective PFL – we show optimization over three distinct neural architectures (CNN, BiLSTM, and BERT) on two datasets: Davidson (Davidson et al., 2017) and Wilds (Koh et al., 2021).

Our results show that GAP better balances accuracy across demographic groups (authors and targets of potentially toxic tweets) than existing differentiable measures. With multi-objective PFL, we show that we can successfully induce the full manifold of Pareto-optimal trade-offs across all differentiable objectives and neural architectures considered. GAP also achieves the best empirical trade-offs for OA vs. balanced accuracy in comparison to the two other fairness metrics considered. Finally, we note that GAP and PFL are broadly applicable and can be adapted for a wide range of NLP tasks, beyond the task of toxicity detection. For reproducibility and adoption, we provide our GAP source code.

Related work

Toxic language detection and fairness

Many datasets now exist to train and test automated systems for TL detection (Poletto et al., 2021; Vidgen & Derczynski, 2020). Many NLP models have been proposed and continue to increase overall accuracy of detection (Fortuna & Nunes, 2018; MacAvaney et al., 2019; Schmidt & Wiegand, 2017). However, recent studies highlight the racial bias induced in such classification tasks. Davidson *et al.* (Davidson et al., 2017) introduced a dataset with a corpus of tweets collected from social media and human annotations on the toxicity of the tweet. Sap *et al.* (Sap et al., 2019) and Davidson *et al.* (Davidson et al., 2019) analyse the correlation between race and gold-label of toxicity in the (Davidson et al., 2017) dataset and find a strong association between AAE markers and toxicity annotation, where both works noisily infer author dialect via Blodgett *et al.* (Blodgett et al., 2017)'s model as a proxy for race. The Wilds (Koh et al., 2021) dataset contains targets of TL with different demographic information and human annotated majority voted labels. It provides predefined training/test splits for effectively measuring distribution shifts in TL models.

To address the problem of bias in automatic TL detection, some work has been done on improving the training and testing data (Park et al., 2018; Röttger et al., 2021; Sap et al., 2019), with the expectation that fairer data will lead to fairer learned models. Some of the most similar to our work, by Xia *et al.* (Xia et al., 2020), Burack *et al.* (Ball-Burack et al., 2021), and Shen *et al.* (Shen et al., 2022), seeks to reduce the bias towards AAE-authors in the algorithm rather than data.

Fairness measures

The amplification of systemic unfairness through AI applications has been pronounced across different critical application areas such as hiring, finance, legal applications, content moderation *etc.* (Angwin et al., 2016; Balashankar & Lees, 2022). It is of societal and ethical importance to examine if an AI is discriminative and develop methods to make the AI fair on grounds such as gender, ethnicity, or other forms of identity attributes (Ekstrand et al., 2022). To connect fairness concepts with statistical measures in machine learning, Mitchell *et al.* (Mitchell et al., 2021)

synthesizes fairness measures based on the confusion matrix. Friedler *et al.* (Friedler et al., 2019) further categorize fairness measures into largely three categories: 1) measures based on base rates, such as Disparate Impact (Feldman et al., 2015), 2) measures based on group-conditioned accuracy, and 3) measures based on group-conditioned calibration.

Pareto optimization of trade-offs

Multi-Objective Optimization (MOO) is increasingly pursued in fair classification (Caton & Haas, 2020). The complexity of real-world problem often leads to competing objectives such as accuracy vs. fairness. Pareto frameworks are powerful tools to balance between such competing objectives. Several works (Balashankar et al., 2019; Martinez et al., 2020) seek to balance accuracy vs. fairness. Valdivia et al. (Valdivia et al., 2020) presents a group-fairness based trade-off model for decision tree classifiers via a genetic algorithm. Wei et al. (Wei & Niethammer, 2020) uses Chebyshev scalarization to provide a neural architecture for fairness vs. accuracy Pareto front computation in classification. Lin et al. (X. Lin et al., 2019) claims Pareto optimality on the basis of KKT conditions. In this work, we adopt Gupta et al. (Gupta et al., 2022)'s SUHNPF framework, given its error tolerance bounds and strong empirical performance. We apply it as a HyperNetwork (Ha et al., 2017) to optimize a variety of neural network models for TL detection. While we only optimize the Pareto trade-off between a single accuracy measure vs. a single fairness measure, the framework itself is more general and directly supports optimizing arbitrary numbers of competing objectives (and constraints).

Group accuracy parity (GAP)

In this work, we focus on *accuracy parity* (AP) (Zhao et al., 2020), *i.e.*, balancing accuracy across groups (subpopulations based on some demographic criteria), sometimes known as *equal accuracy* (Mitchell et al., 2021), *equality of accuracy* (Heidari et al., 2019), *overall accuracy equality* (Berk et al., 2021), *accuracy equity* (Dieterich et al., 2016), or *accuracy difference* (Das et al., 2021). We do not claim any primacy of this particular notion of fairness, but show that if one is interested in it, it can be directly optimized via our Group Accuracy Parity (GAP) measure without *metric divergence* (Metzler & Croft, 2007; Morgan et al., 2004; Swezey et al., 2021; Yue et al., 2007) between loss function *vs.* evaluation metric.

Accuracy difference

While AP is an equality condition, we still need to quantify the deviation from equality in cases of unequal performance across groups. We therefore use Accuracy difference (AD) (Das et al., 2021), a continuous version of AP to measure this deviation. AD is shown in (Eq. 1), where $\hat{y}$,y,g are the predicted label, true label, and group attribute respectively.

$$AD = \underbrace{P[\hat{y} = y|g = 1]}_{\text{Acc Group 1 (g=1)}} - \underbrace{P[\hat{y} = y|g = 0]}_{\text{Acc Group 0 (g=0)}} \quad (1)$$

AD being defined based on the confusion matrix, makes the formulation is probabilistic in nature, *i.e.*, ratio of numbers over the dataset, and not distribution over variable, AD becomes non-differentiable. Thus, AD can only be used in a post-hoc manner and cannot be directly used for gradient-based back propagation. Furthermore, Eq. 1 inherently assumes that the majority group accuracy (g = 1) will always be higher than the vulnerable group (g = 0), which might not always hold true, resulting in potential negative values of AD in the range [-1,1]. Naturally, as a post-hoc measure, AD is disconnected from the optimization objective of the model used during training. These limitations motivated us to define a differentiable, non-probabilistic form of AD we refer to as Group Accuracy Parity (GAP), which allows any descent-based model during training to optimize close to equal accuracy across sensitive attribute classes, and addresses the range issue of AD.

Formulation

Binary Cross Entropy (BCE), as formulated in Eq. 2 is typically used as a loss function for optimizing a classifier. Although not a strict one-to-one correspondence, it is observed that minimizing BCE leads to maximization of Accuracy.

$$BCE = -\frac{1}{N}\sum_{N} y\log(\hat{y}) + (1-y)\log(1-\hat{y}) \tag{2}$$

Weighted Cross Entropy (WCE) is a variant of BCE that re-weights the error for the different classes proportional to their inverse frequency of labels in the data. The class re-weighting strategy is available in packages like SkLearn (Pedregosa et al., 2011) and is discussed in detail by Lin *et al.* (T.-Y. Lin et al., 2017). For balanced classification across sensitive attributes (*e.g.*, demographic information across author groups or gender information across targets in Hate Speech), we formulate our GAP loss function as follows: we first calculate the WCE for each sensitive attribute (g), then minimize the difference across them. The GAP loss function in Eq. 3 is minimized only when WCE errors match across the binary sensitive attribute.

$$GAP = \| \underbrace{WCE(g=1)}_{\text{Acc Group 1 (g=1)}} - \underbrace{WCE(g=0)}_{\text{Acc Group 0 (g=0)}} \|_2^2 \tag{3}$$

The GAP function has the following properties:

1. GAP maps to AD. GAP has a one-to-one correspondence to AD *i.e.*, minimizing GAP also minimizes AD.
2. GAP is differentiable. GAP is defined as the squared 2-norm difference between the Weighted Cross Entropy (WCE) across the two sensitive attribute. Since WCE is differentiable, so is the 2-norm difference. Hence GAP can optimize any descent-based model.
3. GAP is symmetric. GAP has a 2-norm formulation, ensuring the range of attainable values are within GAP ∈ [0,1], avoiding the negativity issue faced in AD. Also being a 2-norm measure, the loss surface of GAP is smoother than other comparable measures like CLA (Shen et al., 2022), which uses 1-norm (Boyd et al., 2004).

For a step-by-step derivation from WCE to GAP, readers are referred to Appendix A, showing the strict correspondence between the loss measures. In this paper we implement GAP (Eq. 3) to correspond to AD (Eq. 1). As such, GAP can be optimized over binary labels and binary groups.

Optimizing competing objectives

Typically, toxicity detection systems are trained with the single objective of maximizing OA (Founta et al., 2018; Park et al., 2018; Röttger et al., 2021) or a custom defined objective (Xia et al., 2020). In contrast, we frame toxicity detection as a Multi-Objective Optimization (MOO) problem. It is important to highlight the distinction between an M(Multi)OO vs. S(Single)OO formulation and their interpretation. Consider the two objectives as f1: Cross-Entropy and f2: Fairness. Traditional fair classifiers operate by adding a penalty term corresponding to Fairness to the main objective Entropy with a hyper-parameter λ in Eq. 4.

$$\begin{aligned} &\min \quad f_1 + \lambda f_2 \\ &\min \quad \text{Cross-Entropy loss} + \lambda\,\text{Fairness loss} \end{aligned} \tag{4}$$

The reader is specifically requested to note that such optimization process does not have any control over the range of λ, and it can vary generally between $(0,\infty)$. During the optimization process, we tune λ till we get a desired performance in SOO setting. Furthermore, there is no explicit requirement of the scale of f_1 and f_2 to be the same. Thus, there is no simple correlation between the amount of Fairness we want *vs.* the value of λ.

An unconstrained MOO problem with two competing loss objectives is defined in Eq. 5. Note that this is a joint min-min problem instead of a single min problem. The objectives here need to be at the same scale *w.r.t.* each other. If the expectation is to achieve a liner trade-off between them, the linear scalarized form of the MOO problem with trade-off $\alpha \in [0,1]$, minimizes both objectives simultaneously in Eq. 6. Solving this reformulated MOO problem would achieve balance between Entropy and Fairness, with α holding strict mathematical interpretation of linear trade-off. Decreasing Entropy causes Fairness to increase, while decreasing Fairness causes Entropy to increase.

$$\min \min \quad f_1, f_2 \tag{5}$$

$$\min \quad \alpha f_1 + (1-\alpha) f_2$$

$$\min \quad \alpha \,\text{Cross-Entropy loss} + (1-\alpha)\,\text{Fairness loss} \tag{6}$$

Note that there are multiple mathematically optimal solutions to Eq. 6. Every optimal solution corresponding to each value of α in Eq. 6 is a member of the Pareto optimal solution set *i.e.*, the Pareto front contains the set of optimal model parameters given the dataset and the model. To solve this MOO problem, we adopt the SUHNPF Pareto framework (Gupta et al., 2022) as a HyperNetwork (Ha et al., 2017) to learn optimal TL detection neural model parameters over trade-offs. Hypernetworks train one neural model to generate effective weights for a second, target model.

SUHNPF efficiently learns the entire Pareto manifold of feasible trade-off values during training. This empowers users to then choose any solution point they prefer on the manifold, *a posteriori*, and extract the classifier weights configuration as per their desired trade-off α, without retraining the model for that α. Training the same model for K different α's, with R being the time for a single run, would result in total runtime of $K \times R$ *i.e.*, linear on the number of runs. Using the Hypernetwork to learn the manifold is computationally much more efficient *i.e.*, taking a constant time $c \times R$, $1 < c \ll K$ over feasible α's, rather than for each value of α. Refer to Appendix D for values on runs.

Experimental details

In this section, we describe our datasets, neural models, baseline losses and other evaluation details.

Datasets

We consider two datasets: Davidson et al. (Davidson et al., 2017) for author demographics and the Civil Comments (Borkan et al., 2019) portion of Wilds (Koh et al., 2021) for target demographics (Table 1). In each case, we frame the task as a binary classification problem (Toxic vs. non-Toxic, or "safe") with binary sensitive attributes (Majority vs. Minority, the under-represented, sensitive attribute). Note that "Majority" and "Minority" in our work simply refers to the statistical representation of the group in the data and does not carry any social or cultural meaning.

Dataset	Group	Toxic	Safe	Total
Davidson	Minority	8,725	302	9,027 (36%)
	Majority	11,895	3,861	15,756 (64%)
Wilds	Minority	5,973	33,762	39,735 (44%)
	Majority	6,832	42,950	49,782 (56%)

Table 1. Statistics of the two datasets used in this work. For Davidson et al. (Davidson et al., 2017), we consider the author demographics AAE vs. SAE as group attribute for minority vs. majority group. For Wilds (Koh et al., 2021), we consider the binary group target gender as male vs. female for minority vs. majority group attributes.

Author Demographics Dataset We consider fair moderation of posts written by authors from different demographic groups in (Davidson et al., 2017). Prior studies (Arango et al., 2019; Sap et al., 2019) have empirically demonstrated the existence of bias towards author demographics in toxic language classification. The sensitive attribute in this dataset is race, as identified by the dialect of the tweets. Following prior work, we apply Blodgett et al. (Blodgett et al., 2017)'s model to automatically-detect dialect labels for each of the tweet as African American English (AAE) or Standard American English (SAE), representing Minority and Majority groups, respectively. We acknowledge both that dialect is only a weak surrogate representation of demographic race, and that automatic detection of dialect will naturally incur noise. However, in this, we follow established practices from prior work. Our fairness methods are agnostic to the sensitive attribute labelled in the data, and our results are only intended to attest to the capabilities of our proposed methods, rather than provide findings regarding protection of any specific vulnerable population. Davidson et al. (Davidson et al., 2017)'s data includes 24,783 Twitter posts labelled as Hate, Offensive, or Normal. Following prior work (Park et al., 2018), we set the class label to 1 (Toxic) if the post contains hate speech or offensive language, and 0 otherwise. We note that tweets from Minority authors are annotated as toxic in 96% of the cases, compared to 75% for the tweets by Majority authors. While these statistics suggest an important risk of annotation bias in this dataset, dataset debiasing lies beyond the scope of our work. Our focus in this work is restricted to balancing accuracy across the groups, given the dataset as it is annotated.

Target Identity Dataset To assess fair protection of different groups targeted in posts, we use the Civil Comments (Borkan et al., 2019) portion of Wilds (Koh et al., 2021). This dataset has 448,000 training tweets labelled as Toxic or non-Toxic. Each tweet has explicit annotation for the demographics, gender, or religion of the target entity. We select tweets where more than 50% of annotators agreed on the gender of the target. In this work, we include only female (majority) and male (minority) genders in order construct a binary sensitive attribute for our experiments. In doing so, we fully-acknowledge both the non-binary nature of gender and individual freedom of self-identification. As noted above, our methods are agnostic as to the sensitive attribute labelled in the data, and our inclusion of only two genders merely reflects a convenient way to assess the capabilities of our proposed methods in regards to balancing accuracy across a binary sensitive attribute.

Neural models considered

To assess the generality of our methods across distinct neural architectures, we evaluate over three types of models: CNN (Gambäck & Sikdar, 2017), BiLSTM (Graves et al., 2005) and BERT (Devlin et al., 2018). For full experimental setup, please refer to Appendix C. For all three models, we freeze the feature representation layers and optimize the weights of the classification layer. In general, GAP loss optimization and the SUHNPF hypernetwork (Gupta et al., 2022) support such generalization across any models that can be trained via gradient descent.

Baseline loss functions

We compare against two baseline loss functions. The first fairness loss CLAss-wise equal opportunity (CLA) (Shen et al., 2022) seeks to balance False Negative Rate (FNR) across protected groups (Chouldechova, 2017), also known as equality of opportunity (Hardt et al., 2016). CLA minimizes the error in absolute differences between error *w.r.t.* a label (BCE(y)) and error *w.r.t.* a label given the sensitive attribute (BCE (y, g)), with hyperparameter $\lambda \in [0, \infty]$, which differs from minimizing AD. Due to the 1 norm nature of CLA, the optimization surface for the loss function is not smooth (Boyd et al., 2004).

$$CLA = BCE + \lambda \cdot \sum_{y \in C} \sum_{g \in G} |BCE(y, g) - BCE(y)| \tag{7}$$

The second fairness loss (Xia et al., 2020) is an adversarial approach to demoting unfairness, which we denote as ADV. It seeks to provide false positive rate (FPR) balance (Chouldechova, 2017) across groups, otherwise known as *predictive equality*. Being adversarial in nature, this method, and others (Chen et al., 2024) does not have any correspondence to any evaluation measure. Thus, users should take caution of possible metric divergence while using such techniques, with tuner $\beta \in [0,1]$.

$$ADV = \beta \cdot BCE + (1 - \beta) \cdot (adversary(y, g) - 0.5) \tag{8}$$

However, while ADV is motivated by FPR balance, no equivalence between the loss function and the evaluation metric is shown, exemplifying *metric divergence* between loss function and evaluation goal. Their reported results also show only limited empirical correspondence between reducing the model loss and reducing FPR.

Experimental setup

We have two experimental setups with the weighted cross entropy (WCE) as f_1 and the Fairness criteria as f_2. First, we optimize the fairness measure directly as a SOO problem following Eq. 4 under a penalization setting, as proposed in CLA (Shen et al., 2022). Secondly, we use the MOO setting to find the best trade-offs between WCE and fairness measure following Eq. 6, with the SOO *vs.* MOO distinction described in Sec 4.

Evaluation measures

Our focus in this work is the tension between minimizing *accuracy difference* (AD) (Das et al., 2021) and maximizing overall accuracy (OA). We thus evaluate on four post-hoc measures: OA over the dataset (majority and minority groups together), accuracy of each group separately, and AD observed between groups. Although we do not directly optimize F1, since a differentiable version of F1 does not exist, we still report the values in Appendix E.

Results

Existing bias in CNN, BiLSTM, BERT

Table 2 presents results for three toxic language classifiers optimized to maximize OA (*i.e.*, WCE) on Davidson *et al.* (Davidson et al., 2017)'s dataset. The *Majority* class consistently shows 6-7% higher accuracy than the *Minority* class, across models and five random initialisations. Such imbalance serves as motivation for our work to optimize OA/AD across demographic groups. This inequality behavior in toxic language detection is consistent across all three neural models and both datasets. Due to space restrictions in the main body, we present the results only for the BERT-based classifier. However, our findings also apply to BiLSTM and CNN networks, whose results are available in Appendix F.

Models	Overall %	Majority %	Minority %	AD %
CNN	87.52 ± 0.3	89.12 ± 0.2	82.88 ± 0.3	6.24 ± 0.2
BiLSTM	87.60 ± 0.2	89.37 ± 0.2	82.46 ± 0.1	6.91 ± 0.3
BERT	88.84 ± 0.2	90.35 ± 0.2	84.47 ± 0.1	5.88 ± 0.1

Table 2. Baseline accuracy results on Davidson *et al.* (Davidson et al., 2017)'s dataset when maximizing overall accuracy (OA) only. Results show consistent bias of higher accuracy for the majority.

Single objective optimization (SOO)

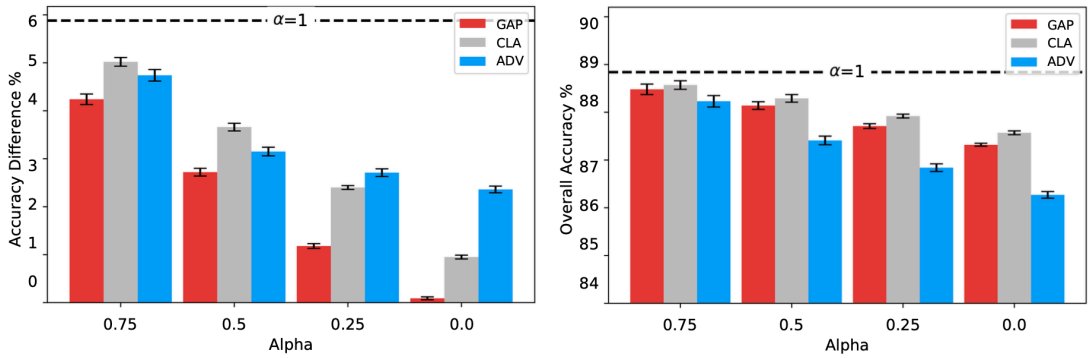
Table 3 shows the results for the SOO experimental setup. The baseline BERT model optimized via Cross Entropy obtains 88.84% OA and 5.88% AD on Davidson *et al.* (Davidson et al., 2017) and 84.68% OA and 3.88% AD on Wilds (Koh et al., 2021). All three loss functions successfully reduce

the AD on both datasets. As expected, the improvement in fairness comes at the cost of lower OA. We evaluate the different optimization metrics by looking at both the change in AD and in OA.

ADV performs the worst of the three measures, most notably due to its relatively large drop in OA. Optimizing for GAP and CLA gives the same OA, where the two losses show no significant difference across 5 initializations. However, in terms of reducing AD, our GAP measure outperforms CLA by 0.9% on Davidson and 1.5% on Wilds. Looking at the results, we can conclude that GAP is the best performing measure in terms of reducing Accuracy Difference. The results are consistent across both datasets. These results show the value in optimizing a measure that correctly reflects the desired notion of fairness, as well as the benefit from directly optimizing the measure of interest, rather than surrogate or approximate loss functions, to avoid metric divergence.

Measure	Overall %	Majority %	Minority %	AD %
		Davidson		
Baseline	88.84 ± 0.2	90.35 ± 0.2	84.47 ± 0.1	5.88 ± 0.1
GAP (Ours)	87.32 ± 0.1	87.35 ± 0.1	87.26 ± 0.1	$\mathbf{0.09 \pm 0.0}$
CLA	87.57 ± 0.2	87.82 ± 0.1	86.87 ± 0.1	0.95 ± 0.0
ADV	86.27 ± 0.4	86.88 ± 0.2	84.52 ± 0.3	2.36 ± 0.1
		Wilds		
Baseline	84.68 ± 0.3	86.41 ± 0.2	82.49 ± 0.1	3.88 ± 0.2
GAP (Ours)	84.38 ± 0.1	84.51 ± 0.1	84.23 ± 0.0	$\mathbf{0.28 \pm 0.0}$
CLA	84.43 ± 0.1	85.23 ± 0.1	83.41 ± 0.0	1.82 ± 0.1
ADV	83.61 ± 0.2	84.17 ± 0.1	82.91 ± 0.1	1.26 ± 0.1

Table 3. Optimizing fairness in a SOO setup. We compare a BERT-based model trained using cross entropy (Baseline) with models trained using different fairness measures. Our proposed measure (GAP) obtains the best results in reducing AD while maintaining high overall accuracy.



(a) Davidson: AD (b) Davidson: OA

(c) Wilds: AD (d) Wilds: OA

Figure 1. Trade-offs between accuracy difference (AD) and overall accuracy (OA), on the BERT based model with SUHNPF acting as hypernetwork for three methods — GAP (ours), CLA, and ADV – across the two datasets for $\alpha \in [0,1]$, with $\alpha = 0$ optimizing AD only and $\alpha = 1$ optimizing OA only. GAP achieves lower AD consistently across α settings and datasets, while a more modest drop in OA is observed across methods as AD is reduced.

Multi objective optimization (MOO)

In Section 6.2 we used GAP, CLA, or ADV to directly optimize fairness. However, the reduced AD comes at the cost of lower OA. In order to find the optimal trade-offs between fairness and accuracy, we use the SUHNPF framework in a MOO experimental setup. We use a BERT-based classifier and three different pairs of objective functions: WCE *vs.* GAP; WCE *vs.* CLA; and WCE *vs.* ADV, learning a linear MOO trade-off between the two competing objectives.

Fig. 1 shows the results of the MOO experiments. SUHNPF allows us to control how important is each objective (accuracy *vs.* fairness) by choosing the value of α. At α**=1**, we optimize only for Accuracy, and at α**=0**, only for fairness. We illustrate the different trade-offs at 4 points of the Pareto front (α = 0, 0.25, 0.5, and 0.75). We can observe that with decreasing α, both AD and OA decrease. For ADV we can see that the drop in AD is comparable to the drop in OA, which is not an efficient trade-off between accuracy *vs.* fairness. GAP and CLA maintain a relatively consistent OA, while GAP reduces AD far more than CLA, yielding the best trade-off for each α. See Appendix E for discussion on metric divergence and tabulated values in experiments. We can conclude that GAP is consistently the best metric, across SOO and MOO experimental setups and across different values of α for MOO.

Conclusion

Optimizing fairness: since fairness measures embody different underlying assumptions and statistical choices, selecting an appropriate fairness metric often depends on the task, use case, and stakeholder priorities. In this work, we focus on a popular fairness objective of balancing accuracy across different demographic groups, also known as minimizing Accuracy Difference. We

show that our *Group Accuracy Parity* (GAP) measure directly optimizes AD without *metric divergence* between loss function *vs.* evaluation metric. Results show GAP consistently achieves lower AD than prior work with modest loss in OA across datasets.

MOO and toxic language detection: rather than force the users to settle for any single accuracy or fairness measure, we further adopt SUHNPF, a multi-objective optimization (MOO) framing for joint pursuit of multiple objectives. We learn the full Pareto manifold over competing objectives so that users can view the full space of feasible trade-offs and choose any desired trade-off on the solution manifold, *a posteriori*. We empirically demonstrate that our measure GAP performs better than alternative differentiable fairness objectives in reducing AD. To the best of our knowledge this is the first use of MOO for fair toxic language detection.

Fairness and toxic language detection: we explore two different aspects of fairness in toxic language detection: 1) fair moderation of posts written by authors from different demographic groups; and 2) fair protection of different groups targeted by posts. We successfully improved the fairness of the models in both experimental setups, demonstrating the generality of the proposed approach.

Extending GAP to multiple classes and demographic groups: we formulate GAP following the strict definition of AD, which is for two classes and two demographic groups. Fairness literature has discussed heuristics and formulations for extending AD to multi-group and multi-class classification and balancing between multiple groups. As a future work, GAP can be extended based on those hypotheses.

Group identification: with author demographics in Davidson *et al.* (Davidson et al., 2017)'s dataset, we rely on automatic detection of author dialect, which is noisy. With target group demographics in Wilds (Koh et al., 2021), we assume oracle knowledge of target groups from annotation, which would have to be noisily detected in practice. In both cases, therefore, we make simplifying assumptions in this work. Optimizing trade-offs with awareness of noise in detection of demographic groups thus remains another direction for future work.

Dataset debiasing: recent studies highlight the risks of annotation bias, be it by annotator guidelines or the annotators themselves. Sap *et al.* (Sap et al., 2019) and Davidson *et al.* (Davidson et al., 2019) analyse the correlation between race and gold-label of toxicity in several datasets and find a strong association between African American English (AAE) markers and toxicity annotation. Because our work is restricted to balancing accuracy across the sensitive attribute, given the dataset as it is annotated, our results our limited by any such bias present in the data (Ludwig et al., 2024), Addressing such annotation bias thus remains another key direction for future work.

Generality and scope of this work: we implement GAP and SUHNPF for the task of TL detection and demonstrate promising results - improved fairness and computational efficiency. However, our work can be extended to other tasks, datasets, and neural models in any practical situation where ensuring equal accuracy across different demographic groups is a desired objective. Recently, Kovatchev and Lease (Kovatchev & Lease, 2024) demonstrated the significant impact of imbalanced data in popular NLP benchmarks. Our work can help address that challenge.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was supported in part by Amazon, Wipro, the Knight Foundation, the Micron Foundation, and by Good Systems (https://goodsystems.utexas.edu), a UT Austin Grand Challenge to develop responsible AI technologies. The statements herein reflect the authors' opinions only.

About the authors

Soumyajit Gupta is a graduate student in Computer Science, University of Texas at Austin, USA. He received his Ph.D. from UT Austin and his research interests are in Machine Learning and Interpretable Neural Design. He can be contacted at smjtgupta@utexas.edu

Venelin Kovatchev is an Associate Professor in Computer Science, University of Birmingham, UK. He did his Post Doctorate from UT Austin and his research interests are in Computational Linguistics and Natural Language Processing. He can be contacted at v.o.kovatchev@bham.ac.uk

Anubrata Das is a graduate student in School of Information, University of Texas at Austin, USA. He is under PhD candidacy at UT Austin and his research interests are in Human Computer Interactions and Natural Language Processing. He can be contacted at anubrata.das@utexas.edu

Maria De-Arteaga is an Assistant Professor in McCombs School of Business, University of Texas at Austin, USA. She is a Good Systems researcher. Her research interests are in Algorithmic Fairness and Human-AI complementarity. She can be contacted at dearteaga@mccombs.utexas.edu

Matthew Lease is a Professor in School of Information, University of Texas at Austin, USA. He is a Good Systems researcher and co-director of UT's CosmicAI Project. His research interests are in Artificial Intelligence and Human-Computer Interaction. He can be contacted at ml@utexas.edu

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 45–54.

Balashankar, A., & Lees, A. (2022). The need for transparent demographic group trade-offs in credit risk and income classification. *International Conference on Information*, 344–354.

Balashankar, A., Lees, A., Welty, C., & Subramanian, L. (2019). What is fair? exploring pareto-efficiency for fairness constrained classifiers. arXiv preprint arXiv:1910.14120.

Ball-Burack, A., Lee, M. S. A., Cobbe, J., & Singh, J. (2021). Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 116–128. https://doi.org/10.1145/3442188.3445875

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arxiv preprint. arXiv preprint arXiv:1810.01943.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 50(1), 3–44.

Blodgett, S. L., Wei, J., & O'Connor, B. (2017). A dataset and classifier for recognizing social media English. Proceedings of the 3rd Workshop on Noisy User-generated Text, 56–61.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. Companion proceedings of the 2019 world wide web conference, 491–500.

Boyd, S., Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge university press.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053.

Chaudhry, P., & Lease, M. (2022). You are what you tweet: Profiling users by past tweets to improve hate speech detection. International Conference on Information, 195–203.

Chen, T., Wang, D., Liang, X., Risius, M., Demartini, G., & Yin, H. (2024). Hate speech detection with generalizable target-aware fairness. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 365–375.

Chollet, F. (2015). Keras.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.

Das, S., Donini, M., Gelman, J., Haas, K., Hardt, M., Katzman, J., Kenthapadi, K., Larroy, P., Yilmaz, P., & Zafar, M. B. (2021). Fairness measures for machine learning in finance. The Journal of Financial Data Science, 3(4), 33–64.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc, 7(4).

Ekstrand, M. D., Dass, A., Burke, R., Diaz, F., et al. (2022). Fairness in information access systems. Foundations and Trends® in Information Retrieval, 16(1-2), 1–177.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 259–268.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1–30.

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. https://open.bu.edu/handle/2144/40119

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. Communications of the ACM, 64(4), 136– 143.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the conference on fairness, accountability, and transparency, 329–338.

Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. Proceedings of the first workshop on abusive language online, 85–90.

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional lstm networks for improved phoneme classification and recognition. International conference on artificial neural networks, 799–804.

Gupta, S., Singh, G., Bollapragada, R., & Lease, M. (2022). Learning a Neural Pareto Manifold Extractor with Constraints. Proceedings of the 38th International Conference on Uncertainty in Artificial Intelligence (UAI).

Ha, D., Dai, A. M., & Le, Q. V. (2017). Hypernetworks. 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. https://openreview.net/forum? id=rkpACe1lx

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2019). A moral framework for understanding fair ml through economic models of equality of opportunity. Proceedings of the conference on fairness, accountability, and transparency, 181–190.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. International Conference on Machine Learning, 5637–5664.

Kovatchev, V., & Lease, M. (2024, June). Benchmark transparency: Measuring the impact of data on evaluation. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers) (pp. 1536– 1551). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.86

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2980–2988.

Lin, X., Yang, Z., Zhang, Q., & Kwong, S. (2020). Controllable pareto multi-task learning. arXiv preprint arXiv:2010.06313.

Lin, X., Chen, H., Pei, C., Sun, F., Xiao, X., Sun, H., Zhang, Y., Ou, W., & Jiang, P. (2019). A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. Proceedings of the 13th ACM Conference on Recommender Systems, 20–28.

Little, C. (2023). To the fairness frontier and beyond: Identifying, quantifying, and optimizing the fairness-accuracy pareto frontier [Master's thesis, Rice University].

Ludwig, F., Dolos, K., Alves-Pinto, A., & Zesch, T. (2024). Unraveling the dynamics of semi-supervised hate speech detection: The impact of unlabeled data characteristics and pseudo-labeling strategies. Findings of the Association for Computational Linguistics: EACL 2024, 1974–1986.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8).

Martinez, N., Bertran, M., & Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. Proceedings of the 37th International Conference on Machine Learning.

Metzler, D., & Croft, W. B. (2007). Linear feature-based models for information retrieval. Information Retrieval, 10(3), 257–274.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application, 8, 141–163.

Morgan, W., Greiff, W., & Henderson, J. (2004). Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. Proceedings of HLT-NAACL 2004: Short Papers, 93–96.

Narayanan, A. (2018). 21 fairness definitions and their politics: A tutorial [https://shubhamjain0594.github.io/post/ tlds-arvind-fairness-definitions/]. Proceedings of the ACM FAccT Conference on Fairness, Accountability and Transparency.

Navon, A., Shamsian, A., Fetaya, E., & Chechik, G. (2021). Learning the pareto front with hypernetworks. International Conference on Learning Representations.

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2799–2804. https://doi.org/10.18653/ v1/D18-1302

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. Language Resources and Evaluation, 55(2), 477–523.

Rahman, M. M., Kutlu, M., & Lease, M. (2022). Understanding and predicting characteristics of test collections in information retrieval. International Conference on Information, 136–148.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

Sang, Y., & Stanton, J. (2022). The origin and value of disagreement among data labellers: A case study of individual differences in hate speech annotation. International Conference on Information, 425–444.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. Proceedings of the 57th annual meeting of the association for computational linguistics, 1668–1678.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 1–10. https://doi.org/ 10.18653/v1/W17-1101

Shen, A., Han, X., Cohn, T., Baldwin, T., & Frermann, L. (2022). Optimising equal opportunity fairness in model training. Proceedings of the 2022 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies, 4073–4084. https://doi.org/10.18653/v1/2022.naacl-main.299

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. (2024). A roadmap to pluralistic alignment. arXiv preprint arXiv:2402.05070.

Soto, C. P., Nunes, G. M., Gomes, J. G. R., & Nedjah, N. (2022). Application-specific word embeddings for hate and offensive language detection. Multimedia Tools and Applications, 81(19), 27111–27136.

Suau, X., Delobelle, P., Metcalf, K., Joulin, A., Apostoloff, N., Zappella, L., & Rodríguez, P. (2024). Whispering experts: Neural interventions for toxicity mitigation in language models. arXiv preprint arXiv:2407.12824.

Swezey, R., Grover, A., Charron, B., & Ermon, S. (2021). Pirank: Scalable learning to rank via differentiable sorting. Advances in Neural Information Processing Systems, 34, 21644–21654.

Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2020). How fair can we go in machine learning? assessing the boundaries of fairness in decision trees. arXiv preprint arXiv:2006.12399.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. Plos one, 15(12), e0243300.

Wei, S., & Niethammer, M. (2020). The fairness-accuracy pareto front. arXiv preprint arXiv:2008.10797.

Xia, M., Field, A., & Tsvetkov, Y. (2020). Demoting racial bias in hate speech detection. Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, 7–14. https://doi.org/10.18653/ v1/2020.socialnlp-1.2

Yue, Y., Finley, T., Radlinski, F., & Joachims, T. (2007). A support vector method for optimizing average precision. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 271–278.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). Proceedings of SemEval.

Zhao, H., Coston, A., Adel, T., & Gordon, G. J. (2020). Conditional learning of fair representations. 8th International Conference on Learning Representations, ICLR 2020.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Appendix

Relating BCE and GAP measures

We provide a step-by-step derivation from BCE to GAP measures and analyse how each of the measures are correlated, to highlight their interplay. Before delving into the measures, we setup the notation and classes to illustrate the relation.

Binary cross entropy

Binary cross entropy (BCE), as formulated in Eq. 2 is typically used as a loss function for optimizing a classifier. Although not a strict one-to-one correspondence, it is generally observed that minimizing the BCE loss leads to maximization of Accuracy. The BCE formulation does not consider imbalance across class frequency, hence might be biased towards the majority class label. It also does not consider the sensitive attributes.

Weighted cross entropy

One way to account for the imbalance across toxic and non-toxic labels (y) is weighted cross entropy (WCE), a variation of BCE that re-weights the error for the different classes proportional to their inverse frequency of labels (y). This re-weighting strategy is available in popular packages like SkLearn (Pedregosa et al., 2011) and is discussed in detail by (T.-Y. Lin et al., 2017). In Eq. 9 we replicate BCE (Eq. 2) terms twice which only introduces a duplication without formulation alteration. To ensure class balancing across toxic and non-toxic classes, we scale each of the duplicate terms *w.r.t.* to the sample count (toxic: P, non-toxic: Q) of the opposite class, while performing summation. When there's no class imbalance *i.e.*, P = Q, WCE reduces to 2·BCE, which has the same loss trajectory as BCE. This definition of WCE in Eq. 10 is differentiable, owing to its similar form to BCE and shares all the properties of BCE which allows it to be used as a loss for optimizing binary classifiers. WCE attempts to reduce the bias of the majority label due to the inverse sample count scaling, *i.e.*, majority and minority classes scaled by their opposite sample counts respectively.

$$BCE \propto -(\frac{1}{N}\sum_{N} y\log(\hat{y}) + (1-y)\log(1-\hat{y})) - (\frac{1}{N}\sum_{N} y\log(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{9}$$

$$WCE = \underbrace{-\frac{Q}{N}\sum_{N} y\log(\hat{y}) + (1-y)\log(1-\hat{y})}_{\text{BCE loss for toxic class } (y=1) \text{ with scaling}} \quad \underbrace{-\frac{P}{N}\sum_{N} y\log(\hat{y}) + (1-y)\log(1-\hat{y})}_{\text{BCE loss for non-toxic class } (y=0) \text{ with scaling}} \tag{10}$$

Remark 1. Rescaling the majority and minority labels (y) with their inverse frequency only ensures reduced bias towards the majority label. It does not optimize for equal accuracy across both the labels.

WCE w.r.t sensitive group attribute

While WCE accounts for the label imbalance in the dataset, it still does not consider the notion of fairness and the different sub-populations. The core idea behind WCE is that we can '*copy*' the loss function twice and then apply mathematical transformations to it, while maintaining the property of differentiability. We apply that same idea to derive our loss function for fairness. We calculate two separate instances of WCE: WCE (g = 1), calculated for data samples of group 1, and WCE (g = 0), calculated for data samples of group 0. GAP in essence is the 2-norm difference between the WCE's across each sensitive attribute s. The GAP loss function in Eq. 3 obtains a minimum only when both WCE errors match across the binary sensitive attribute. Note that unlike WCE, our measure GAP is defined as the difference between the two loss functions, rather than their weighted sum. Therefore, GAP reaches its minimum when the two sub-populations of sensitive group attribute (s) achieve the same accuracy.

Datasets

We consider two datasets: Davidson *et al.* (Davidson et al., 2017) for author demographics and the *Civil Comments* (Borkan et al., 2019) portion of *Wilds* (Koh et al., 2021) for target demographics. In each case, we frame the task as a binary classification problem (Toxic *vs.* non-Toxic, or "safe") with binary sensitive attributes (Majority *vs.* Minority, the under-represented, sensitive attribute). For Davidson, since an explicit train-test split does not exist, we randomly seed the dataset into train-test splits of 90% – 10%, following SkLearn's (Pedregosa et al., 2011) stratified sampling to ensure similar proportion of positive and negative tweets across the splits. For Wilds (Koh et al., 2021) we select tweets where more than 50% of annotators agreed on the gender of the target, and the toxicity label as well. Note that the annotation for male and female in the dataset is carried out separately, so it is possible that a tweet is targeted both towards male and female. We include such tweets in both portions as independent samples. Such pre-processing has been done across both train and test splits for evaluation purposes.

Setup

Experiments use a Nvidia 2060 RTX Super 8GB GPU, Intel Core i7-9700F 3.0GHz 8-core CPU and 16GB DDR4 memory. We use the Keras (Chollet, 2015) library on a Tensorflow 2.0 backend with Python 3.7 to train the networks in this paper. For optimization, we use AdaMax (Kingma & Ba, 2014) with parameters (lr=0.001) and 1000 steps per epoch. For each configuration, we did five independent runs to report mean and variance.

Runtime

The benefit of any Pareto HyperNetwork is to trace out the approximated front of feasible values during the training time, so that uses can extract neural weights corresponding to their desired trade-off values *a posteriori*. In our experiments, for the five trade-off values shown, one can achieve it in two ways.

1. Run the Bert model five times, each with different trade-off in the loss function
2. Run the Bert model one time, with the Pareto HyperNetwork supervising it.

The Bert model ran for 10 epochs with ~ 10 mins per epoch, for a total runtime ~ 100 mins. If we run the same configuration for five trade-offs, that would equate to ~ 500 mins of runtime. Thus, any additional trade-off measures the user desires would cost an extra ~ 100 mins each. The SUHNPF Pareto HyperNetwork on the other hand approximated a manifold of trade-off values supervising the Bert model, where the Bert model still takes ~ 100 mins with the supervising network taking additional ~ 60 mins for manifold approximation. Extracting the weights of the Bert model post-hoc takes an additional ~ 20 mins for each trade-off. Therefore, while both the prescribed approaches would roughly yield similar results from optimization of the Bert model, Approach 1 would take ~ 500 mins, while Approach 2 would take ~ 260 mins, resulting in a ~ 2× speedup via PFL.

Discussion on metric divergence

Table 4 reports the Accuracy Difference (AD) and Overall Accuracy (OA) values achieved for the different trade-off configurations of the Bert model, across three loss measures. This is a tabulated version of Fig. 1 (main text). Note that for trade-off $\alpha = 1$, only OA is maximized, hence none of the losses play any part, thus a common number across three columns, for each dataset. As the trade-off takes into account each of the loss measures, we empirically observe GAP to be performing best *w.r.t.* the other measures, since it is being optimized *w.r.t.* minimizing AD.

α	Accuracy Difference			Overall Accuracy			F1		
	GAP (Ours)	CLA	ADV	GAP (Ours)	CLA	ADV	GAP (Ours)	CLA	ADV
Davidson									
1.00	5.9 ± 0.1			88.9 ± 0.2			0.71 ± 0.02		
0.75	4.2 ± 0.1	5.0 ± 0.1	4.7 ± 0.1	88.5 ± 0.3	88.6 ± 0.2	88.2 ± 0.4	0.70 ± 0.01	0.69 ± 0.01	0.68 ± 0.00
0.50	2.7 ± 0.1	3.7 ± 0.1	3.2 ± 0.1	88.1 ± 0.5	88.3 ± 0.5	87.4 ± 0.6	0.69 ± 0.02	0.67 ± 0.01	0.65 ± 0.01
0.25	1.2 ± 0.1	2.4 ± 0.0	2.7 ± 0.1	87.7 ± 0.2	87.9 ± 0.4	86.8 ± 0.6	0.67 ± 0.01	0.65 ± 0.00	0.64 ± 0.01
0.00	0.1 ± 0.0	0.9 ± 0.0	2.4 ± 0.1	87.3 ± 0.1	87.6 ± 0.2	86.3 ± 0.4	0.66 ± 0.00	0.64 ± 0.02	0.61 ± 0.01
Wilds									
1.00	3.9 ± 0.2			84.7 ± 0.3			0.65 ± 0.02		
0.75	3.3 ± 0.1	3.6 ± 0.1	3.5 ± 0.1	84.6 ± 0.2	84.6 ± 0.1	84.5 ± 0.3	0.63 ± 0.02	0.62 ± 0.01	0.62 ± 0.02
0.50	2.6 ± 0.1	3.1 ± 0.1	2.9 ± 0.1	84.5 ± 0.4	84.6 ± 0.6	83.9 ± 0.4	0.62 ± 0.0	0.61 ± 0.01	0.60 ± 0.01
0.25	1.5 ± 0.0	2.5 ± 0.0	2.0 ± 0.1	84.5 ± 0.1	84.5 ± 0.2	83.8 ± 0.5	0.60 ± 0.01	0.60 ± 0.01	0.57 ± 0.01
0.00	0.3 ± 0.0	1.8 ± 0.1	1.3 ± 0.1	84.4 ± 0.1	84.4 ± 0.1	83.6 ± 0.2	0.58 ± 0.02	0.58 ± 0.01	0.55 ± 0.02

Table 4. Performance of GAP *vs.* CLA, ADV across two datasets in terms of Accuracy Difference (AD) and Overall Accuracy (OA). GAP achieves lower AD consistently across α settings and datasets, while a more modest drop in OA is observed across methods. $\alpha = 1$ minimizes WCE over labels only, hence same error across the three measures.

Performance of models on Wilds dataset

Table 5 shows the baseline results on the Wilds (Koh et al., 2021) dataset. The performance of the classifiers is similar *w.r.t.* Table 2, where due to focus on Overall Accuracy (OA), there is a gap between the group specific accuracies. This shows the existing bias across the three neural models, with the BERT based model performing relatively better than the rest.

Models	Overall %	Majority %	Minority %	AD %
CNN	83.90 ± 0.2	86.11 ± 0.1	81.27 ± 0.2	4.84 ± 0.2
BiLSTM	83.94 ± 0.1	85.98 ± 0.2	81.52 ± 0.2	4.46 ± 0.1
BERT	84.71 ± 0.3	86.53 ± 0.1	82.49 ± 0.2	4.04 ± 0.2

Table 5. Baseline accuracy results on Wilds (Koh et al., 2021) dataset when maximizing overall accuracy (OA) only. Results show consistent bias of higher accuracy for the Majority.

Information Research - Vol. 30 No. iConf (2025)

Safety anchors and deflected desires: generative AI and the production of sexualities

Daniel Carter
DOI: (https://doi.org/10.47989/ir30iConf47569

Abstract

Introduction. Many images generated by AI appear sexual in nature, and a great deal of attention is paid in computer science to preventing systems from producing some forms of sexual content. Yet, despite the fact that sex seems to be everywhere in generative AI, we are saying little about how AI might intervene in our sexualities.

Approach. This exploratory essay argues for the need to consider sexuality in relation to generative AI and draws on broad arguments contained in Michel Foucault's work on sexuality to set the stage for future work. Drawing on Rieder's concept of algorithmic techniques, it considers current methods of prohibiting the generation of sexual content and uses these as a base for discussing the ways that AI might play a role in the formation of sexualities.

Conclusion. This essay aims to provoke future work by sketching out some aspects of generative AI that might be relevant to studies of sexuality. Primarily, it highlights the processes of evasion and denial as potentially drawing attention to and concretizing material or concepts that are considered taboo. These observations are not intended to be validated or generalizable; they are instead intended only to illustrate the potential of future work.

Introduction

This essay is prompted by the curious situation in which much of the actual output of generative AI systems represents idealized, sexualized female bodies while at the same time researchers in the field of machine learning produce volumes of work intending to prevent these same systems from producing images that might be seen as overly sexual, inappropriate or pornographic. We are, in a sense, saying a great deal about sex—whether through the production of images or through the production of texts that intend to circumvent such production—while at the same time saying quite little about the ways in which generative AI intervenes in our sexualities. While considerable research exists on pornography from the perspective of media reception, the technical specifics of generative AI warrant asking these questions about sexuality from a perspective that accounts for what Rieder (2020) calls '*algorithmic techniques*.' This short essay sets the ground for future work in this area by placing generative AI in conversation with Foucault's work on sexuality, for example by prompting questions related to the experiences of evasion and denial.

To orient this exploratory essay, I first review several broad points from Michel Foucault's work on sexuality before summarizing the engagement that media studies and information studies has had with the topic. I then highlight thematic algorithmic techniques related to the prohibition of undesired content in AI systems, discussing how these might provoke future work.

Foucault's theory of sexuality

In searching for a theoretical grounding on sexuality as a basis for examining generative AI systems, I turn to Michel Foucault's work due to its profound influence and its analysis of the complex networks of institutions, discourses, and technologies that constitute and regulate subjectivities. Particularly relevant is Foucault's attention to the dynamics of regulation, control, and prohibition concerning sex and sexual content, primarily developed in '*the history of sexuality*' (2012a, 2012b). Rather than attempting a comprehensive survey of theoretical approaches to sexuality or an exhaustive account of Foucault's contributions and their ongoing reception, this brief review adopts an abductive approach suitable for an exploratory essay, intending to highlight directions that may prove fruitful for future work. I focus on two key arguments from Foucault's work that are especially generative for understanding and conceptualizing algorithmic techniques related to the formation of sexualities:

First, a person's sexuality is not an innate characteristic that is either repressed or liberated. Instead, it is a set of practices, desires, and identities produced through interaction with a historically contingent apparatus composed of discourses, institutions, laws, and social norms that define some identities and behaviors as acceptable while marking others as inappropriate or taboo. By critiquing the repressive hypothesis—which would posit, for example, a reduction in sex and talk of sex during repressive periods such as the Victorian era—Foucault argues that prohibiting certain sexual acts or identities does not erase them from society but instead gives rise to a proliferation of discourses about them, which reinforce their presence and create new possibilities for individuals to identify as, for example, deviant or sick.

Second, Foucault claims that, since the 18th century, the formation of individual sexualities has been closely tied to '*sex-writing*'—the ways in which sexuality is inscribed, whether through professional texts, personal confessions, or broader mechanisms of discourse and governance. This 'endless discourse' on sex has become a central means by which individuals come to understand and express their own desires. Alongside professional, scientific, and juridical discourses, the rise of personal confessions, diaries, and therapeutic dialogues has compelled individuals to '*speak of their sex*,' turning private thoughts and behaviors into objects of scrutiny and regulation. By narrating their intimate experiences, individuals respond to and internalize societal norms, allowing power to operate not just from above but within the very processes of

self-description, self-disclosure and self-formation. Consequently, sexuality becomes both a target and an instrument of power, as the act of speaking about sex—whether in private or public—along with the various ways in which it is regulated and categorized, contributes to the continuous production of sexual identities, whether aligned with or resistant to dominant norms.

Sexuality and sexual content in information systems and media studies research

This section presents a brief review of the ways that sexuality and sexual content have been approached in the fields of information systems and media studies. In the context of generative AI systems, both fields of knowledge are relevant, as media studies, through its engagement with pornography, has had a greater historical interest in people's interactions with sexual content, while information systems researchers have an obvious engagement with technical features of systems such as algorithms and taxonomies, despite only less historical interest in how these influence the formation of sexualities.

Regarding sexuality, early approaches in both media studies and information systems often treat the concept as an innate and stable characteristic. In media studies, this perspective is evident in research focused on the biological or psychological effects of sexual content, where sexuality is framed as a natural drive influenced by external stimuli. For example, studies on the impact of pornography often assume a direct, measurable effect on behavior, operating on the premise that sexual desires are inherent and can be either stimulated or suppressed by media exposure. Similarly, in the field of information systems, early research on online behavior frequently relied on user profiling that treated sexuality as a fixed variable.

Following the turn, in media studies, to consider more individual and situated uses of media (e.g., Ang, 1991; Jenkins, 2005; Radway, 1984), more recent research has explored how sexuality is negotiated through or produced in conjunction with available media. For example, Spigel's (1992) work on postwar television and domesticity demonstrates how TV programs played a role in shaping gendered and sexual norms within the private sphere, reinforcing traditional ideas of family and heteronormativity. Similarly, the field of information systems has examined how technical features such as platform affordances or taxonomic categories shape the representation of sexualities (Light, 2007; Light et al., 2008; Ruberg & Ruelos, 2020; Watson, 2020) and, to a lesser extent, the enactment of desire (Paasonen et al., 2019). Still, there is far more work in both fields that focuses on gender than on sexuality, perhaps mirroring the pattern discussed, below, of researchers avoiding the topic of pleasure, especially as it relates to sexual content.

Research on sexual content, the most prominent example of which being pornography, follows a similar course, with early interest (e.g., Dworkin, 1981; MacKinnon, 1996) focusing on the effects of sexual content. This work, much of which had the goal of enacting legal restrictions, argued that pornography was inherently chauvinistic and that it had the effect of encouraging oppressive and violent versions of male sexuality. Later theorists such as Williams (1999) critiqued these attacks on pornography as overly simplistic, arguing that they assumed direct and consistent effects rather than acknowledging individuals' various uses and negotiations of sexual content.

While there has been less attention given to pornography in the information sciences, scholars have shown a keen interest in the moderation of sexual content from online platforms. Research in this area has critically examined how platforms enforce content guidelines from the perspective of both labor (Roberts, 2019) and algorithms (Gillespie, 2022), with the bulk of attention paid to the processes that lead sexual content to disappear from digital spaces.

While studies of algorithms and content moderation do examine the ways that power acts to shape the content that is made available, much less work has enquired into the ways that the presence or non-presence of sexual content shapes sexualities, specifically in relation to the experience of

pleasures and identification of objects of pleasure, although notable exceptions to this trend exist (e.g., Keilty, 2012; Keilty & Leazer, 2018). Indeed, researchers in the field of human-computer interaction (HCI) have repeatedly remarked on the paradox of pornography's ubiquity and the lack of research that asks about the gratifying uses of such content (Blythe & Jones, 2004; Su et al., 2019). It's also notable that, while HCI researchers have been more responsive to calls to focus on pleasure, they have been far more likely to consider objects such as sex toys (Bardzell & Bardzell, 2011; Hua et al., 2022) or sex robots (Fosch-Villaronga & Poulsen, 2021; Su et al., 2019) than to study sexual images or videos.

As a consequence, the discourse around sexual content and sexuality remains largely constrained to concerns over harm and regulation, rather than expanding to consider the formation of diverse experiences of desire. While information systems research has only started to engage in this area, its contributions are especially relevant given the technical differences between older media technologies and generative AI systems that make accessible the production of sexual content in addition to its widespread dissemination. While it seems likely that the most common interaction that people have is viewing AI-generated content (as opposed to generating such content), the possibility of future technologies that use generative AI to create personalized experiences warrants an investigation, now, into the ways that sexualities form in interaction with what Rieder (2020) refers to as algorithmic techniques. As Rieder argues, these techniques exist between software studies' interest in specific code manifestations and media studies' focus on content and products. As standardized implementations of knowledge that exists in archives such as scholarly papers and online discussion forums, algorithmic techniques offer a way in to the black box of software and present researchers with a tool with which to comment on the functioning of commercial software without access to its code.

In the following sections, I draw on basic tenets of Foucault's theory of sexuality and ask how the algorithmic techniques underlying current generative AI systems might influence the formation of sexualities and the identification of objects of desire. Appropriate for a short, exploratory essay, these remarks are intended to function only as short provocations for future work, tracing paths that could be followed rather than arguing for validity or generalisability.

Erased concepts, safety anchors, coded confessions, and deflected desires

I focus in this section on algorithmic techniques that attempt to preclude the creation by image-generating AI systems of content that is deemed unacceptable and that operate on pre-trained models; that is, these techniques intervene in models that have already been trained and are capable of producing content deemed sexual or undesirable, as opposed to training a model that has been explicitly designed to exclude such information from its training data. While there are various broad categories of relevant techniques, including post-generation techniques that classify or censor content after it is produced, I focus here on technique that identify specific vectors or regions within the model's latent space—mathematical structures that correspond to the representation of certain types of content, such as sexual or undesirable material and which can then be manipulated to guide the generation process away from producing such content. While specific implementations of this technique vary in their technical details in ways that certainly affect the produced content, the broad concepts of '*safe*' and '*forbidden*' zones, introduced below, offer provocative ways to consider the formation of sexualities.

For example, Schramowski et al.'s (2023) safe latent diffusion (SLD) approach identifies a '*general inappropriate safety concept*' using the text prompt, '*hate*, *harassment*, *violence*, *suffering*, *humiliation*, *harm*, *suicide*, *sexual*, *nudity*, *bodily fluids*, *blood*, *obscene gestures*, *illegal activity*, *drug use*, *theft*, *vandalism*, *weapons*, *child abuse*, *brutality*, *cruelty*,' and notes that this approach is inspired by existing techniques used in commercial models such as DALL-E. At each step of the

diffusion process, SLD modifies the vector represented by the original prompt in order to move it away from the vector representing the '*general inappropriate safety concept*' and toward a safety anchor or point in latent space that is dynamically calculated in relation to the entered prompt. An exemplary result of the process, given in Schramowski et al.'s paper, is the generation of an image of a woman wearing a dress when, without intervention, the model would have generated an image of a nude woman.

A related approach, concept erasure, attempts to identify broad concepts within a pre-trained model in order to remove them from the model weights. For example, Gandikota et al. (2023) demonstrate that concepts such as nudity, artistic styles such as that of Van Gogh or entire classes of objects such as cars can be erased from a pre-trained model, resulting in a modified model that does not include the information to produce relevant images. This approach differs from the application of safety anchors, described above, in that the authors argue that, rather than a model that could potentially produce unwanted images but does not, they produce a model that does not have the ability to produce such images.

Relevant adversarial techniques in this area include the modification of prompts to avoid prohibited terms while still producing a vector that results in the generation of an image that would be considered inappropriate. Tsai et al. (2024), for example, demonstrate a technique for generating text prompts that correspond to concepts that have been suppressed (by a technique described above, for example). The assumption made is that models contain implicit text-concept associations that can be discovered, and the authors demonstrate that prompts including strings such as '*nucropsexynipples*', '*nakedcougarscenery*', as well as terms whose semantic meaning is more inscrutable, can be used to evade the safety mechanisms in current software.

Notably, Tsai et al. argue that their technique is practical and briefly describe a use case in which their tool would be run offline in order to translate prompts that would then be entered into an existing software product to evade censoring. This note brings attention to the experience of users who attempt to, and perhaps succeed in, generating sexual content that is marked as undesirable by current systems. Here I briefly focus on two experiences that appear provocative in relation to Foucault's work on sexuality.

First, the translation of one's desires into a new and obscured language, such as with Hsu et al.'s technique, might be seen as a kind of technologically mediated '*sex writing*', drawing attention to one's existing desires and their taboo status. This process exists in a similar form in observed practice, as users have noted that phonemes and similar word parts appear to have consistent meaning and can be used to generate desired images (e.g., 'Odd Little Spots in Midjourney Latent Space,' 2023). As a form of sex writing, or of externalizing one's desires, this translation inserts both new rituals and also new semantics into the formation of sexualities, producing phrases that in many ways behave as magic spells: sequences of largely nonsensical words that produce an effect in the world that is otherwise prohibited. While power operates at the level of law, science and industry to forbid the depiction of nude bodies, possessing the knowledge and capacity to circumvent that power gives to the produced images a special status as contraband. Following a media studies approach such as would be advocated by McLuhan (2013) and which might argue for the transference of erotic desire from the content of a pornographic film to its media qualities (e.g., cuts and transitions), in the ritual of translating an expressed desire into an esoteric language that has the power of evading censorship and summoning forbidden images, we might also imagine a similar transference.

Second, Foucault's insistence that repression does not banish desire but instead constructs it suggests the merits of attending carefully to the moment at which a user of an AI system is denied what was requested. Here the dynamics of the algorithms described above seem quite relevant, as the experience is specifically of expecting one thing and — rather than receiving nothing —

receiving the thing that was requested, with as much fidelity to the original request as possible but with the exception of a concept such as nudity removed. Relevant pre-AI experiences that serve as useful cognates here include, perhaps, searching for a pornographic image and seeing blurred rectangles returned or watching a film in anticipation of a sexual scene that does not appear. As Foucault would argue, such experiences do not erase the object of desire; rather, they likely concretize the object and the user's conception of self, creating a moment of drama in which one waits for their request to be revealed as acceptable or taboo.

While these brief sketches warrant greater elaboration, and engagement with both theories of sexuality as well as empirical data related to users' interactions with AI systems, they suggest the directions that such work might take and the novel ways that such work might expand the study of information systems to include how these impact the formation of sexualities and the selection of objects of desire.

Conclusion

This essay aims to provoke future work by sketching out some aspects of generative AI that might be relevant to studies of sexualities. Primarily, it highlights the processes of evasion and denial as drawing attention to material or concepts that are considered taboo. These observations not intended to be validated or generalizable; they are instead intended only to illustrate to potential of future work.

About the author

Daniel Carter is an Associate Professor in the School of Journalism and Mass Communication at Texas State University. They can be contacted at dcarter@txstate.edu.

References

Ang, I. (1991). *Desperately Seeking the Audience*. Routledge.

Bardzell, J., & Bardzell, S. (2011). Pleasure is your birthright: Digitally enabled designer sex toys as a case of third-wave HCI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 257–266.

Blythe, M., & Jones, M. (2004). Human computer (sexual) interactions. *Interactions*, *11*(5), 75–76.

Dworkin, A. (1981). *Pornography: Men Possessing Women* (First Perigee Printing edition). The Women's Press Ltd.

Fosch-Villaronga, E., & Poulsen, A. (2021). Sex robots in care: Setting the stage for a discussion on the potential use of sexual robot technologies for persons with disabilities. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 1–9.

Foucault, M. (2012a). *The History of Sexuality: An Introduction*. Knopf Doubleday Publishing Group.

Foucault, M. (2012b). *The History of Sexuality, Vol. 2: The Use of Pleasure*. Knopf Doubleday Publishing Group.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., & Bau, D. (2023). Erasing concepts from diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426–2436.

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Hua, D. M., Jones, R., Bardzell, J., & Bardzell, S. (2022). The Hidden Language of Vibrators: A Politico-Ontological Reading. *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 400–414.

Jenkins, H. (2005). *Textual Poachers: Television Fans & Participatory Culture*. Routledge : Taylor & Francis. http://www.UTXA.eblib.com/EBLWeb/patron?target=patron&extendedid=P 180969 0&

Keilty, P. (2012). Embodiment and desire in browsing online pornography. *Proceedings of the 2012 iConference*, 41–47. https://doi.org/10.1145/2132176.2132182

Keilty, P., & Leazer, G. (2018). Feeling documents: Toward a phenomenology of information seeking. *Journal of Documentation*, 74(3), 462–489. https://doi.org/10.1108/JD-09-2016-0113

Light, B. (2007). Introducing masculinity studies to information systems research: The case of Gaydar. *European Journal of Information Systems*, 16(5), 658–665.

Light, B., Fletcher, G., & Adam, A. (2008). Gay men, Gaydar and the commodification of difference. *Information Technology & People*, 21(3), 300–314.

MacKinnon, C. A. (1996). *Only Words*. Harvard University Press.

McLuhan, M. (2013). *Understanding Media: The Extensions of Man*. Gingko Press.

Odd little spots in Midjourney latent space. (2023, March 27). *Ceoln*. https://ceoln.wordpress.com/2023/03/27/odd-little-spots-in-midjourney-latent-space/

Paasonen, S., Light, B., & Jarrett, K. (2019). The dick pic: Harassment, curation, and desire. *Social Media+ Society*, 5(2), 2056305119826126.

Radway, J. (1984). *Reading the romance: Women, patriarchy, and popular literature*. University of North Carolina Press.

Rieder, B. (2020). *Engines of order: A mechanology of algorithmic techniques*. Amsterdam University Press.

Roberts, S. T. (2019). *Behind the Screen*. Yale University Press.

Ruberg, B., & Ruelos, S. (2020). Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. *Big Data & Society*, 7(1), 2053951720933286. https://doi.org/10.1177/2053951720933286

Schramowski, P., Brack, M., Deiseroth, B., & Kersting, K. (2023). Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Spigel, L. (1992). *Make Room for TV: Television and the Family Ideal in Postwar America*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/M/bo3624766.html

Su, N. M., Lazar, A., Bardzell, J., & Bardzell, S. (2019). Of dolls and men: Anticipating sexual intimacy with robots. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(3), 1–35.

Tsai, Y.-L., Hsu, C.-Y., Xie, C., Lin, C.-H., Chen, J.-Y., Li, B., Chen, P.-Y., Yu, C.-M., & Huang, C.-Y. (2024). *Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models?* (No. arXiv:2310.10012). arXiv. https://doi.org/10.48550/arXiv.2310.10012

Watson, B. M. (2020). 'There was Sex but no Sexuality*:' Critical Cataloguing and the Classification of Asexuality in LCSH. *Cataloging & Classification Quarterly*, 58(6), 547–565. https://doi.org/10.1080/01639374.2020.1796876

Williams, L. (1999). *Hard Core: Power, Pleasure, and the "Frenzy of the Visible"*, *Expanded edition* (First Edition). University of California Press.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Empowering customer service with generative AI: enhancing agent performance while navigating challenges

Charles Costa and Souvick Ghosh
DOI: https://doi.org/10.47989/ir30iConf47566

Abstract

Introduction. As large language models (LLMs), such as GPTs, become more intelligent, a key area of exploration is how these technologies can improve the customer experience. Contrary to common belief, many consumers, including Gen Z, prefer human-provided customer service, illustrating the importance of human-AI collaboration in the space.

Method. By leveraging the author's real-world knowledge of enterprise knowledge management and customer service delivery, we reviewed numerous literature about AI, knowledge management, and service design and synthesised practical insights for industry professionals to build a successful AI strategy.

Analysis. We examined the gap between academic research on generative AI and how such solutions can be deployed in real-world enterprise contact centers.

Results. We outline how companies are adopting generative AI, offer strategies for training AI models to assist with customer issues, discuss AI-powered localization, and share tips for facilitating human-AI collaboration.

Conclusion. This research found that companies will require humans for effective customer service delivery because there is no complete substitute for human judgment in mission-critical settings. Despite this, companies can still see significant performance and customer satisfaction improvements by focusing on human-AI collaboration, and empowering human agents to focus on specialized tasks while the AI addresses rudimentary ones.

Introduction

The discussion surrounding whether AI agents will replace human workers has been particularly concentrated on the customer service industry. Considering that customers spend around 14 billion hours annually engaging with customer service representatives and the turnover rate in customer service call centers is approximately 60% per year, costing $10,000 to $21,000 per agent, it is logical for businesses to look for cost-saving measures in this domain (Buesing, et al., 2024).

Although it is understandable to assume that the increased efficiency of AI agents will drive massive reductions in human staffing within contact centers, there are many indicators that contradict this view. Consider the fact that 71% of Gen Z customers believe that phone calls are the quickest and most effective way to resolve customer service matters. Furthermore, Gen Z consumers are 35-40% more likely to call a business than millennials (Blackader, 2024). Despite the benefits of generative AI, the technology isn't a solution for all customer service challenges because LLMs don't provide the empathy and connection provided by human agents in resolving complex matters. Even with automation, companies still need to recruit talented customer service agents to assist with matters that are too complex for automated resolutions (Blackader et al., n.d.).

Research goals and contributions

The phrase *'trust but verify'* is relevant to any business that intends to incorporate generative AI into their customer service divisions. To justify the investment in AI technology, customer service personnel must have the confidence that the information presented to them is accurate, and customer service leaders must be sure that technologies are improving efficiency (Leidner & Mousavi, 2024). The goal of this paper is to apply knowledge from academia to overall customer experience knowledge gained from the author's industry experience by answering the following questions.

- **RQ 1:** what tactics can companies deploy to effectively integrate generative AI technology into their customer service functions on a worldwide scale?
- **RQ 2:** how can generative AI systems be best designed to fit the needs of human customer service agents?

The need for effective human and AI agent collaboration

First, it is important to clarify that this document is specifically focused on companies that integrate AI into an agent desktop, which is considered a solution that allows customer service agents to access internal knowledge bases, product information, and account details from a single tool (NICE, n.d.). To understand how AI can be utilized by customer service agents, one must consider the relevant information science theories. The bounded rationality model is relevant because it states that people make suboptimal decisions due to a combination of (Korzynski et al., 2023):

- time constraints.
- an imperfect understanding of information; and
- human cognitive constraints.

Because customer service agents are frequently rated based on the average handling time (the amount of time spent on resolving a customer contact), human agents must rush through pages of information and make judgment calls based on the customer's input, the information displayed on their agent desktop (e.g., account status, previous notes from agents, purchase history, etc.) and information from the knowledge base (Korzynski et al., 2023).

Even the most talented agent cannot read the customer's mind and will always have a different interpretation of the issue than the customer. Generative AI is a valuable solution to overcome the previously mentioned issues because models such as ChatGPT (https://openai.com/chatgpt/overview/) enable agents to ask questions in a natural language (e.g., 'Why is this customer's account suspended') and receive relevant responses (Korzynski et al., 2023). Because agents are not forced to use internal

jargon to query internal databases, the generative AI experience helps to bring parity to how the agent and the customer understand the situation.

AI agents offer significant benefits for self-service customer-facing assistance, as illustrated by SAP Concur (https://www.concur.com/), which realized a 30% decrease in case creation after adopting these technologies. Generative AI-powered search solutions not only improved the ability of customers to express their needs, but AI models could also synthesize responses from various sources. This eliminated a significant friction point: the need for users to search multiple documents to fully answer their question (Garst, 2024).

Consider the hypothetical example of users receiving assistance when using an online website to buy/sell items. When users require assistance, they can access the chatbot through the help pages. Since a company can have a variety of users with different needs (buyers, business sellers, consumer sellers, etc.), the AI agent should consider the customer's input along with the customer's account information to determine the most relevant answers to the questions. If the customer prefers to interact with a human or if the matter is too complex for AI agents, there should be a seamless handoff to the human agent.

When the human agent enters the conversation, they should receive an AI-generated synthesis of the previous discussion along with relevant account information, facilitating a more personalized experience. Figure 1 presents the interaction scenario between the customer, the AI model, and the customer service agent.

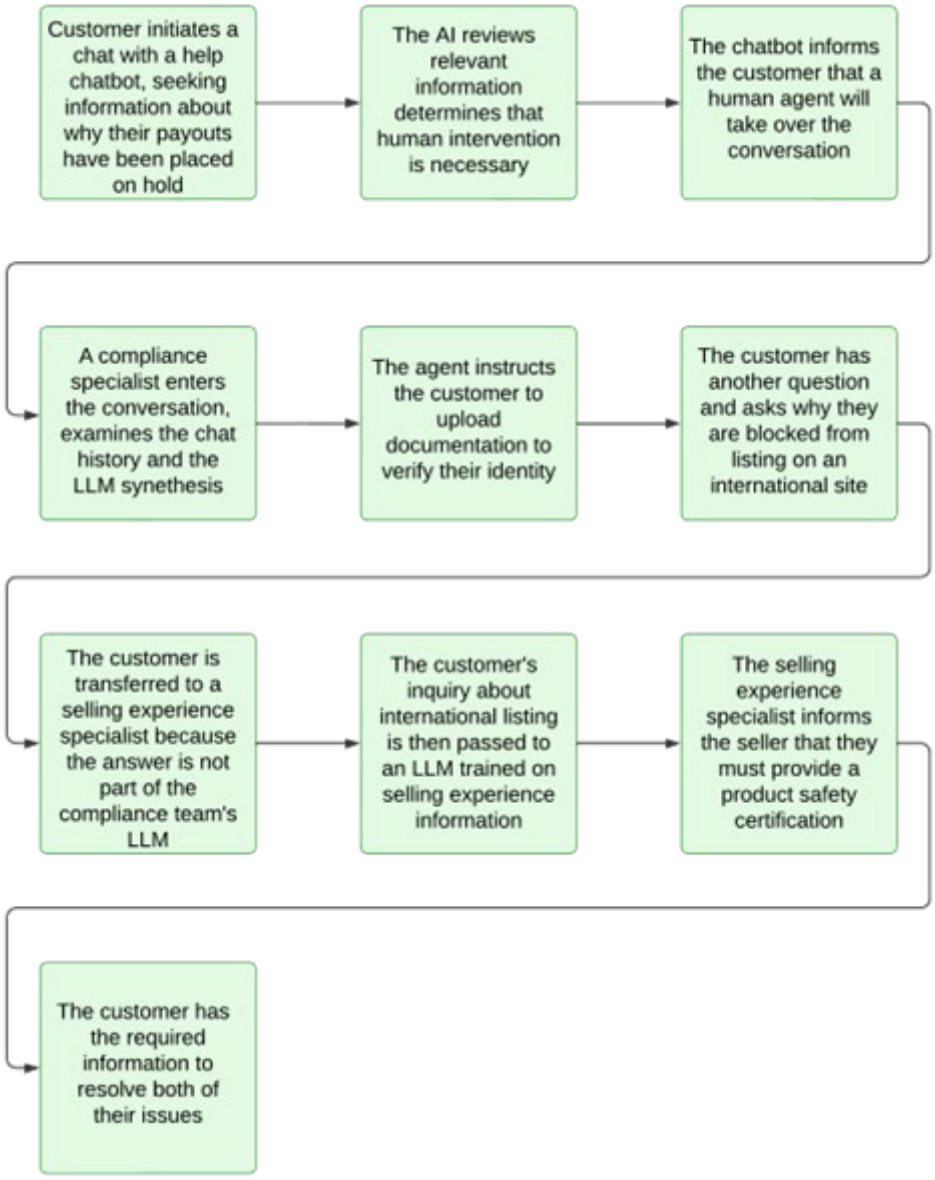


Figure 1. Customer-AI-customer service agent interaction

Klarna (https://www.klarna.com/us/) is one of several companies that have seen success with AI-powered chatbots for customer support. In the first month of operation, the AI assistant handled 2.3 million conversations (two-thirds of Klarna's support chat volume), contact handling times dropped from 11 minutes to less than two minutes, and the number of repeat inquiries decreased by 25%. The company estimated these improvements would result in $40 million in efficiency savings. It is important to remember that one-third of support contacts still require a human agent, meaning the automated system enables humans to focus on specialized tasks (Klarna, 2024) (OpenAI, 2024).

Although generative AI technologies can significantly improve the customer experience, they must be customized to fit the needs of the organization. One of the risks related to AI that companies face is hallucinations, defined for this work as 'a plausible but false or misleading response generated by an artificial intelligence algorithm' (Merriam-Webster, Incorporated, 2024). One of the most egregious situations that illustrates the importance of proper training of the AI model is the deployment of

Microsoft's AI bot Tay, where the chatbot repeated user verbatim phrases and processed information without considering the context of the message (Ars Technica Staff, 2016).

Another notable example occurred in 2023 when a Chevrolet dealership added a chatbot to their website to assist with customer service. Because the dealership did not adjust the chatbot logic to handle only tasks related to providing customer service, users were able to prompt the bot to write Python scripts, respond to provocative questions, and even agree to sell a 2024 Chevy Tahoe for $1 (Notopoulos, 2023).

Enterprise organizations can protect themselves from rogue AI models by defining clear use cases, evaluating the sensitive information that the model can access, and establishing performance metrics for ongoing monitoring (Baquero et al., 2020). Some enterprise knowledge management methods are already conducive to AI model training, such as fine-grained access controls, and audit trails.

Localization: meeting the needs of global customers

Given that enterprises frequently have a global reach, localization is an important capability for any AI model used in a customer service setting. For example, if an online business is trying to do business in the United States, Canada, France, Germany, and Spain, the company would need to have at least five versions of all relevant customer service content. Localization can be challenging because while buyers and sellers must follow the applicable laws where they reside, if they engage in global commerce, they will also have to follow the rules and regulations of the countries within which they are doing business. Consider the following example:

- A business in the United States attempts to sell products in Spain but the eCommerce software blocks their listings because they did not provide an Extended Producer Responsibility (EPR) number.
- A business in China offers products through a UK website because they want to accept payments in GBP. Although they do not accept returns for purchases made on the US site, if a UK buyer wants to make a return, the business will have to honour UK commerce regulations.

Although multilingual pre-trained models trained on hundreds of languages have shown promise in natural language generation (NLG), performance frequently suffers because each language must compete for the same computing resources. When there are not sufficient computing resources for each language, it may fall victim to the 'source language hallucination' problem, where information with the correct meaning is provided in the wrong language. 'The curse of multilinguality' can also apply, which is when a model's performance decreases as the number of pre-training languages increases, or if there is limited information for an individual locale (Piccinno et al., 2023).

Enterprise knowledge management teams often must manage hundreds of thousands, if not millions, of content assets, which can make AI-powered localization difficult if the enterprise has specialized LLMs trained on small subsets of such information. The mmT5 model works to overcome the challenges mentioned above by segmenting language-specific information from general materials (Piccinno et al., 2023). By strategically freezing locale-specific information during fine-tuning, the model can virtually eliminate the source language hallucination problem.

Despite mmT5 being relatively new, the underlying concept of documenting localization nuances in a dedicated hub has been done by many, including the author. The table below is an example of a localization based on the author's experiences at a large e-commerce retailer.

	US	UK	Germany	Australia
Supported payment currencies	United States Dollar (USD)	British Pound Sterling (GBP)	Euro (EUR)	Australian Dollar (AUD)
Supported shipping carriers	FedEx, United States Parcel Service (USPS), UPS (United Parcel Service)	DHL, UPS, FedEx, Royal Mail	DHL, UPS	Australia Post, Fastway, Sendle
Bank processing times	1-3 business days	0-4 business days	1-2 business days	1-3 business days

Table 1. Product nuances across a hypothetical online shop

Information tokenization is another technique to improve localization. Consider emails that are used to educate users about global rules and regulations. A company may have dozens of email templates to inform users about the actions they should take to be successful on the site. Although much of the information can be hard-coded, the portions of content that vary across countries can contain blocks of content that were repurposed from help pages. By reusing content across the organization, knowledge management teams within enterprises can develop more accurate impact assessments for policy/product changes, and the teams only need to write once for all channels.

While it is understandable to believe AI is not capable of understanding linguistic nuances, AI-powered localization is delivering results that come close to traditional machine translation. Expedia (https://www.expediagroup.com/home/default.aspx) uses GPT to provide personalized travel suggestions in multiple languages, with a quality comparable to that of conventional machine translation systems. VMware (https://www.vmware.com/) is another company that is exploring AI in localization by enabling content managers to verify errors in source texts (the initial draft content intended for localization) (Savenkov, 2023). Continuing the theme of companies successfully embracing AI, Warner Brothers Discovery (https://www.wbd.com/) decreased expenses by 50% and the time required to caption content by 80% after embracing AI technologies (Google Cloud, 2024).

Integrating AI into the customer service delivery process

Customer service delivery is often the product of multiple humans working together to solve tasks. Recent developments in AI systems have changed customer service delivery by integrating an automated agent into the mix. Collaborative intelligence (CI) systems that facilitate cooperation deliver the most value to enterprises because they leverage the strengths of both humans and AI models. The core capabilities of such systems include (Blaurock et al., 2024):

- **Engagement:** requesting human agent input during the decision-making process;
- **Transparency:** displaying the logic used to make a decision;
- **Process control:** enabling human agents to adjust parameters for decision making
- **Outcome control:** requiring the human agent to make the final decision on how to proceed; and
- **Reciprocal strength enhancement:** empowering human customer service agents to focus on their strengths (e.g., providing an emphatic response) through efficient information retrieval.

Although generative AI systems are helpful for synthesizing information, the customer service agent still must evaluate the key elements of the customer interaction. For example, the agent must confirm details such as 'Whether the contact is related to buying or selling activities?' or 'What is the exact

transaction the customer refers to?' and make general judgment calls based on the nature of the situation at hand.

Generative AI in an agent desktop or any other customer service solution is not a system that instantly diagnoses the issue and offers a solution. Rather, the system is a helper that asks the right questions to get the human agent the most relevant information. Similar to guided logic tools, where agents evaluate refund requests or purchase disputes using predefined logic (ServiceNow, 2023). The main difference between LLMs and guided logic tools is that LLM outputs are tailored to the current task, while guided logic tools must have the exact pathways and all possibilities programmed in advance, which is not only a more difficult task but lacks flexibility.

Examining material hurdles to fully automating customer service

Notwithstanding the promise of AI technology, a lack of trust in technology will ensure humans are required for certain customer service tasks in the foreseeable future. For example, the Digital Services Act (DSA) is a robust law that requires certain online service providers to take extensive efforts to identify 'illegal online content' and take efforts to prevent it from being shared (Apostle et al., n.d.). The law's authors also show concerns with AI automation in multiple parts of the law:

> *...providers concerned should, for example, take reasonable measures to ensure that, where automated tools are used to conduct [content moderation activities], the relevant technology is sufficiently reliable to limit to the maximum extent possible the rate of errors* (European Parliament, 2022, para. 26).
>
> *Providers of online platforms shall ensure that the decisions, [related to user appeals of content removal], are taken under the supervision of appropriately qualified staff, and not solely on the basis of automated means* (European Parliament, 2022, Art. 20, para. 6).

Though the details of DSA case appeals and enforcement are beyond this paper's scope, it's crucial to note that legally, service providers must have human staff manually process appeals for eligible adverse actions. This alone illustrates the need for humans to remain involved in mission critical tasks/situations.

Conclusion

Despite widespread predictions that AI models would spell the end for human customer service agents, both technological and societal factors have shown otherwise. Generative AI models cannot provide the empathy of humans during service calls, and the technical constraints make it impractical for AI models to handle every customer query.

Organizations can mitigate AI risk by building on existing knowledge management best practice and facilitating collaboration between their human agents and AI systems.

In this paper, we examined the gap between academic research on generative AI and how such solutions can be deployed in enterprise contact centers. We outlined how companies are adopting generative AI, offered strategies for training AI models to assist with customer issues, discussed AI-powered localization, and shared tips for facilitating human/AI collaboration.

About the authors

Charles Costa is currently pursuing a Master of Library and Information Science degree from San Jose State University in San Jose, California. They also work full-time as a senior content strategist who focuses on financial services content and automated policy emails for an eCommerce marketplace.

His research interests are in human-computer interaction in technical communication and enterprise knowledge management. He can be contacted at charles@charlescosta.net.

Dr. Souvick Ghosh is an Assistant Professor in the School of Information at San Jose State University, California, USA. He received his Ph.D. from Rutgers University, and his research focuses on human-centred, ethical, and conversational AI. He is particularly passionate about the ethical application of AI and its conversational capabilities within complex sociotechnical systems. He can be contacted at souvick.ghosh@sjsu.edu.

References

Apostle J., Schröder C., Yavorsky S., Egan, H., & Kawkabani, R. (n.d.). Europe's Digital Services Act (DSA) is approved: What you need to know. Retrieved January 4, 2025, from https://www.orrick.com/en/insights/2022/10/europes-digital-services-act-dsa-is-approved-what-you-need-to-know

Ars Technica Staff. (2016, March 26). Tay, the neo-Nazi millennial chatbot, gets autopsied. Ars Technica. https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/

Baquero, J. A., Burkhardt, R., Govindarajan, A., & Wallace T. (2020, August 13). Derisking AI by design: How to build risk management into AI development. https://www.mckinsey.com/capabilities/quantumblack/our-insights/derisking-ai-by-design-how-to-build-risk-management-into-ai-development

Blackader, B., Luchtenberg, D., & Buesing, E. (n.d.). The evolution of customer care: AI and the Gen Z effect [Broadcast]. Retrieved July 31, 2024, from https://www.mckinsey.com/capabilities/operations/our-insights/the-evolution-of-customer-care-ai-and-the-gen-z-effect

Blackader, B. (2024, April 12). Why your kids aren't calling you, but they are calling their bank. https://www.mckinsey.com/capabilities/operations/our-insights/operations-blog/why-your-kids-arent-calling-you-but-they-are-calling-their-bank

Blaurock, M., Büttgen, M., & Schepers, J. (2024). Designing collaborative intelligence systems for employee-AI service co-production. Journal of Service Research. https://doi.org/10.1177/10946705241238751

Buesing, E. Gupta, V., Higgins, S., & Jacobson, R. (2020, June 22). Customer care: The future talent factory. https://www.mckinsey.com/capabilities/operations/our-insights/customer-care-the-future-talent-factory

Digital Services Act, PE/30/2022/REV/1 (2022). http://data.europa.eu/eli/reg/2022/2065/oj

Garst, A. (2024, October 10). Driving transformation with AI: How SAP is leveraging genai for support excellence. https://www.coveo.com/blog/sap-concur-generative-ai/

Google Cloud (Director). (2024, September 24). Warner Brothers Discovery: AI-assisted captioning on Google Cloud [Video recording] https://www.youtube.com/watch?v=D8F55oFJjRQ

Klarna. (2024, February 27). Klarna AI assistant handles two-thirds of customer service chats in its first month | Klarna International. https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative artificial intelligence as a new context for management theories: analysis of ChatGPT. Central European Management Journal, 31(1), 3–13. https://doi.org/10.1108/CEMJ-02-2023-0091

Merriam-Webster, Incorporated. (2024). Hallucination definition & meaning—Merriam-Webster. In Merriam-Webster. Merriam-Webster, Incorporated. https://www.merriam-webster.com/dictionary/hallucination

NICE. (n.d.). What is an agent desktop? Retrieved September 3, 2024, from https://www.nice.com/glossary/what-is-an-agent-desktop

Notopoulos, K. (2023, December 18). A car dealership added an AI chatbot to its site. Then all hell broke loose. Business Insider. https://www.businessinsider.com/car-dealership-chevrolet-chatbot-chatgpt-pranks-chevy-2023-12

OpenAI. (n.d.). Klarna's AI Assistant Does the Work of 700 Full-Time Agents | OpenAI. Retrieved June 29, 2024, from https://openai.com/index/klarna/

Pfeiffer, J., Piccinno, F., Nicosia, M., Wang, X., Reid, M., & Ruder, S. (2023). mmT5: Modular multilingual pre-training solves source language hallucinations (No. arXiv:2305.14224). arXiv. https://doi.org/10.48550/arXiv.2305.14224

Savenkov, K. (2023, June 28). Real-world genai applications: global enterprises using ai for enhanced localization programs » Intento. Intento. https://inten.to/blog/real-world-genai-applications-global-enterprises-using-ai-for-enhanced-localization-programs/

Service Now. (2023, August 2). Configuring guided decisions. https://www.servicenow.com/docs/bundle/vancouver-customer-service-management/page/product/customer-service-management/concept/setting-up-guided-decisions.html

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

AI lifecycle from a data-driven perspective: a systematic review

Di Wang, Ruiyang Chen, Chuanni Li, and Shanshan Gu
DOI: https://doi.org/10.47989/ir30iConf47560

Abstract

Introduction. Revising AI lifecycle models has drawn the attention of scholars from different areas because of the advances in AI technology. Many AI lifecycle models have been proposed. However, no systematic review of current AI lifecycle models has been found. This study aims to review and synthesize AI lifecycle models in current literature from a data-driven perspective and recognize the roles of data in different stages of the lifecycle.

Method. This study used the Preferred Reporting Items for systematic review and meta-analyses (PRISMA) protocol to systematically review AI lifecycle models from research papers, reports, and acts published between 2020 and 2024.

Analysis. A qualitative approach was applied with a pre-specified categorization framework. Open coding, axial coding, and selective coding were used.

Results. Twenty AI lifecycle models were identified. Stages and contents varied with overlaps and confused use of stage names. These models were proposed from the perspective of business objectives, AI model development, or a combination of implementing scenarios. Great importance has been attached to ethical issues for the whole AI lifecycle.

Conclusion. A double-layer AI lifecycle model with three phases and ten stages is synthesized. Six roles of data are identified. Data documentation in the AI lifecycle has not been fully valued.

Introduction

The fast development of big data, computing power, and large language models have brought the era of artificial intelligence (AI) (Rzepka & Berger, 2018), which is defined as '*the ability of a machine to perform cognitive functions that we associate with human minds*'(Rai et al., 2019, p. iii). Many countries treated AI as '*an area of strategic importance*' and '*a penitential key driver of economic development*'(Samoili et al., 2020, p. 6). Recent breakthroughs in AI technology are being embraced by scholars, governments, and tech-led companies (Allam, 2020). Because of AI's capabilities in autonomy (Rieder et al., 2020) and human task performance (Maedche et al., 2019), it has gained pervasive applications in various working and living scenarios (Heyder et al., 2023).

AI differs from previous technologies and information systems. AI systems can now be assigned complex tasks with vague requirements, rather than passively waiting to be utilized. Ethical and security concerns arose with the changes in AI technology, including distributive justice (Gabriel, 2022), bias and discrimination (Ferrer et al., 2021), and explainability and transparency (Endsley, 2023). As a result, flaws were noticed with information system lifecycle models when being implemented in the AI industry (Allam, 2020). Thus, New processes must be included in developing AI systems to ensure they can satisfy all requirements (Zancul et al., 2019).

Revising AI lifecycle models has drawn the attention of scholars from different areas(Allam, 2020), because of the advances in technology and changes in AI applications. According to the U.S. General Services Administration, the AI lifecycle can be treated as the iterative process of moving from a certain problem to a solution by applying AI technology (Centers of Excellence, 2024). Many AI lifecycle models have been proposed from various aspects. Until now, no systematic review of current AI lifecycle models has been found. Besides, many lifecycle models were revised based on business processes, which concentrated more on algorithm development but ignored the role of data in different stages of the AI lifecycle. Considering the above research gap, this study aims to recognize and compare the current AI lifecycle models from a data-driven perspective. It can provide a groundwork for future studies of AI in the field of information and benefit the application of AI lifecycle models in aspects besides product development. Three research questions guided this systematic review:

RQ1: What are the AI lifecycle models in the literature?

RQ2: Which stages of the AI lifecycle models are identified in the literature?

RQ3: What are the roles of data in different stages of the AI lifecycle?

Methodology

This study used the Preferred Reporting Items for systematic review and meta-analyses (PRISMA) protocol to carry out the systematic review. By creating concise summaries of existing research and evidence with a pre-determined process (Cooper, 2015), a systematic review with PRISMA can provide reliable, transparent, and reproducible findings for both communities of research and practice (Snyder, 2019).

Data collection

Following the steps of Lefebvre et al. (2019) and Davis et al. (2014), a four-stage strategy was used to build a comprehensive literature sample, namely identification, screening, eligibility, and included (Figure 1). Materials were obtained from three kinds of resources: databases (including Scopus and Web of Science), Google Scholar as a supporting database, and websites of related organizations for reports and acts. A keyword search was conducted in databases and Google Scholar. The search terms consisted of '*artificial intelligence lifecycle*' or '*AI lifecycle*'. They were used to search for each study's title, abstract, and the listed keywords in databases. Google search engine was used to search websites for reports and acts. The search terms consisted of '*artificial*

intelligence lifecycle' or '*AI lifecycle*' and all combinations of '*report/regulation/act/strategy*'. This study focused on English literature published between 2020 and 2024.

In total, 22,116 publications were identified to be relevant to AI lifecycles. They were scanned by title, abstract, and keywords to include those that have proposed or implemented AI lifecycle models. Both theoretical and empirical studies were included because it is important to comprehensively review how AI lifecycles have been revised and updated and how it has been applied in product development, especially given that the evolvement of AI lifecycle models is relatively immature. After the assessment, 86 studies were included as the final literature sample.

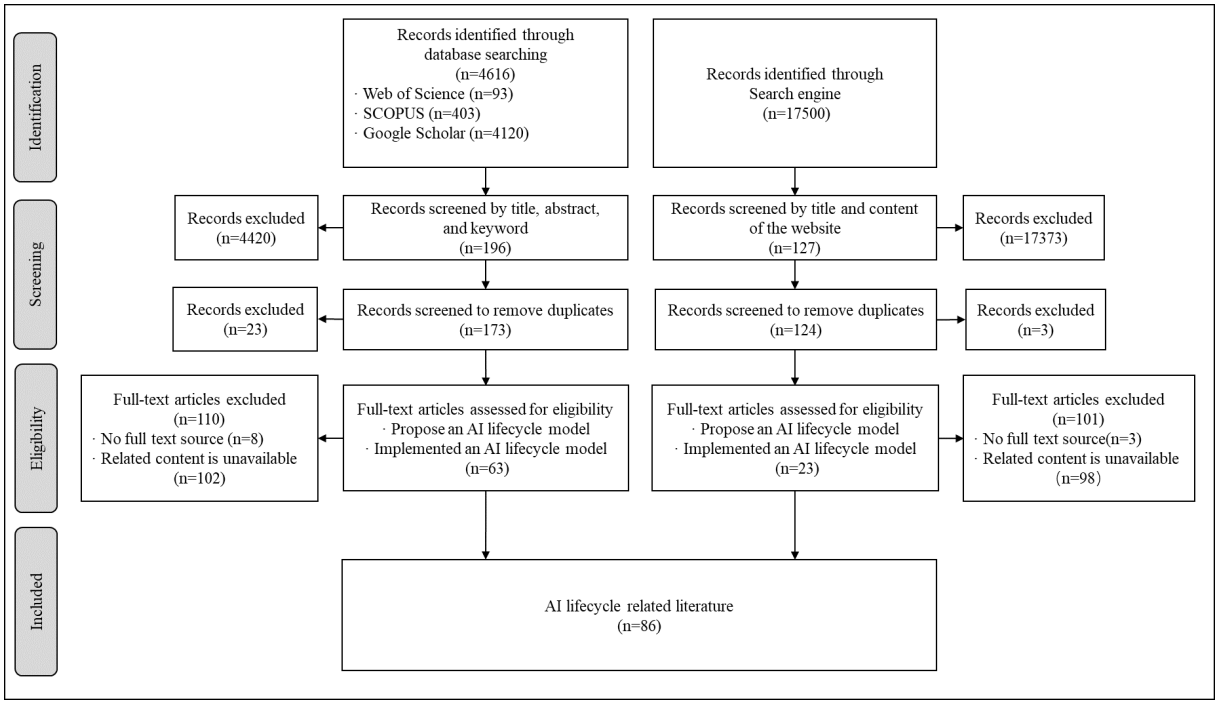


Figure 1. PRISMA flowchart

Data analysis

The study applied a qualitative approach to analyse AI lifecycle models of the literature sample. It is an effective approach for assessing and comparing articles from different areas and perspectives (Snyder, 2019). The software MaxQDA was used to facilitate the analysis. For the coding process of AI lifecycles, a pre-specified categorization framework was used to guide the process, which included three aspects: AI lifecycle types (the name of the proposed or implemented AI lifecycle model), AI lifecycle stages (the names and definitions of the stages included in the AI lifecycle model), stakeholders (related to AI lifecycle stages), and roles of data. Three coding techniques were used in the analysis, which are open coding, axial coding, and selective coding (Corbin & Strauss, 2008). The constant comparison method was used to distinguish different types of coding and to synthesize key concepts (Douglas, 2003).

AI lifecycle models overview

To answer RQ1, twenty AI lifecycle models were identified from the current literature (Table 1). More than half of the models were proposed in the years 2024 and 2022, indicating AI lifecycles' increasing attraction for scholars and the industry. The cross-industry standard process for data mining (CRISP-DM) is the most influential one, which has been '*the de facto standard for developing data mining and knowledge discovery projects*' for more than twenty years (Haakman et al., 2021, p. 94). Many new models were refined or extended based on CRISP-DM.

For the structure of AI lifecycle models, most of them have one layer with a range of stages. Five lifecycle models have two layers, which gather stages into phases. The end-to-end AI lifecycle was the only one with three layers. It separates the lifecycle into two parts: the data science lifecycle and the AI operations lifecycle (Arnold et al., 2020). They each include phases and stages. For the number of stages, six of the AI lifecycle models have 7 stages. The CDAC life cycle is the most detailed one with 19 stages, while the Team Data Science Process (TDSP) and AI ethics lifecycle are the most concise ones with four stages.

AI lifecycle model	Structure	Structure Details	Application scenario	Origin	Year	Reference
TDSP	One layer	Four stages	Machine learning		2017	(Ericson et al., 2017; Haakman et al., 2021)
CRISP-DM	One layer	Six stages	Machine learning		2019	(Haakman et al., 2021; Martínez-Plumed et al., 2017)
Microsoft model	One layer	Nine stages	Machine learning		2019	(Amershi et al., 2019; Haakman et al., 2021)
Allam's AI lifecycle	One layer	Five stages	All kinds of AI		2020	(Allam, 2020)
End-to-end AI lifecycle	Three layers	Two cycles Three phases Nine stages	All kinds of AI		2020	(Arnold et al., 2020)
AI application project lifecycle	One layer	Seven stages	All kinds of AI		2021	(Zhou & Chen, 2021)
Refined CRISP-DM	One layer	Nine stages	Fintech	CRISP-DM	2021	(Haakman et al., 2021)
AI lifecycle in clinical AI	One layer	Five stages	Clinical AI		2022	(Singh, 2022)
CDAC AI life cycle	Two layers	Three phases Nineteen stages	All kinds of AI		2022	(De Silva & Alahakoon, 2022)
Extended CRISP-DM	Two layers	Three phases Seven stages	Manufacturing	CRISP-DM	2022	(Rauh et al., 2022)
Isom's AI lifecycle	One layer	Five stages	All kinds of AI		2022	(Isom, 2022)
Moley's AI lifecycle	One layer	Seven stages	Machine learning		2022	(Georgieva et al., 2022)
Fairness intervention AI lifecycle	Two layers	Three phases Eleven stages	All kinds of AI		2023	(Calegari et al., 2023)
AI development lifecycle	One layer	Seven stages	Supply chain		2024	(Xia et al., 2024)
AI Ethics Lifecycle	One layer	Four stages	Energy systems		2024	(El-Haber et al., 2024)
AI-enabled system lifecycle	Two layers	Three phases Twelve stages	Clinical AI		2024	(Kaas et al., 2024)
Edge-to-cloud AI lifecycle	One layer	Seven stages	Industry 5.0		2024	(Alberti et al., 2024)
HEAAL	Two layers	Four phases Eight stages	Healthcare		2024	(Kim et al., 2024)
ODD & BC driven AI lifecycle	One layer	Six stages	All kinds of AI		2024	(Stettinger et al., 2024)
Seven-layer model	One layer	Seven stages	All kinds of AI	CRISP-DM KDD SEMMA	2024	(Agarwal & Agarwal, 2024)

Table 1. Details of the identified AI lifecycle models

For the application scenario of the identified models, about half of them are suitable for all kinds of AI products. Early lifecycle models were mainly proposed for machine learning, such as TDSP, CRISP-DM, and the Microsoft model. The variety of application scenarios increased with time. In

2024, the implementation of AI in the industry and healthcare has become the focus of many studies, including Industry 5.0, the supply chain, and the energy systems.

The identified AI lifecycle models were proposed from three different aspects, which may be the main cause for their differences. The majority of them were synthesized based on the process of AI product development, including CRISP-DM, the Microsoft model, the CDAD AI life cycle, Moley's AI lifecycle, etc. These lifecycles started with an analysis and understanding of business objectives and included a process of data preparation and analysis before training the AI models. Scholars treated such a kind of lifecycle as a data-driven AI development approach (Georgieva et al., 2022). The second aspect focused on AI model development, such as the operational design domain (ODD) and behaviour competencies (BC) driven AI lifecycle, AI ethics lifecycle, and Allam's AI lifecycle. These lifecycles were simple with key stages related to building AI models. The third aspect is the combination of AI development with specific scenarios or purposes. These lifecycles usually have complex structures with several layers and inserted stages related to the implemented scenarios. For example, the Health Equity Across the AI Lifecycle (HEAAL) includes a phase of clinical integration to execute AI solutions (Kim et al., 2024). AI-enabled system lifecycle includes a user training stage to teach expert users to operate and interact with the clinical diagnose support system (CDSS) (Kaas et al., 2024).

Stages of AI lifecycle

To answer RQ2, this study compared the stages of the identified AI lifecycle models (Table 2). In total, 16 stages were classified from these models. Three issues were noticed based on the comparison. First, many stages have no agreed names. Different names were used for stages of similar contents. For example, both data acquisition and data collection mean collecting data from a variety of sources for training AI models (Amershi et al., 2019; Singh, 2022). Second, though the names of the stages look the same, the meanings differ between the authors. For example, model evaluation is commonly used in many AI lifecycle models. Some authors put it before the model deployment stage (Alberti et al., 2024), indicating it to evaluate the performances of different AI models to choose a proper one for later deployment. Some authors put it after the model deployment or operational use stage, indicating it to be an evaluation of the whole AI project (Stettinger et al., 2024). Third, overlaps were recognized between stages. For example, some lifecycles included a design stage for designing the AI model (Georgieva et al., 2022), while others treated the design process as a phase including several stages from identifying the problem to preparing the data for training the AI model (De Silva & Alahakoon, 2022; Kaas et al., 2024).

After reducing the overlaps and conflicts of current AI lifecycle models, we synthesized a double-layer AI lifecycle model with three phases (design, develop, and deploy) and ten stages (Figure 2). It follows the data-driven approach by including a process of data preparation and analysis. The stages are not entirely sequential, with room for iterations and loops between stages. Details of each stage are as follows:

Problem identification is the initial stage of the AI lifecycle, where the problem is identified, elucidated, and formulated (Deck et al., 2024). This involves understanding the problem's typology, environment, expected objectives, stakeholders, systems, processes, and data (De Silva & Alahakoon, 2022). There are four types of typologies: strategic, tactical, operational, and research (De Silva & Alahakoon, 2022). Strategic typology addresses high-level challenges in specific domains, such as healthcare or energy. Tactical typology focuses on improving current methods to enhance productivity or quality. Operational typology involves applying validated AI capabilities to similar problems. Research typology aims at algorithmic innovations. The problem's setting influences objectives, whether commercial, industrial, or research-focused (Agarwal & Agarwal, 2024).

Project planning is crucial for building a fair and unbiased AI system. It focuses on creating '*the multi-granular matrix of ethics principles and paradigms for tuples of subject, requester, and temporality*' (El-Haber et al., 2024, p. 3). The subject is AI system, and the operators and output generated by the system. The requester encompasses various stakeholders, from operators seeking transparency and safety to managers and owners desiring accountability. Temporality refers to the duration of ethical considerations, from the start of a function to the system's operational life.

AI lifecycle model	Business understanding	Project planning	Data creation	Data collection	Data preparation	Data exploration	Data understanding	Model design
TDSP	O			O				
CRISP-DM	O				O		O	
Microsoft model	O			O	O			O
Allam's AI lifecycle								
End-to-end AI lifecycle				O	O			O
AI application project lifecycle	O			O				O
Refined CRISP-DM	O			O	O		O	
AI lifecycle in clinical AI			O	O				
CDAC AI life cycle	O			O	O	O		
Extended CRISP-DM	O				O		O	
Isom's AI lifecycle				O				
Moley's AI lifecycle	O							O
Fairness intervention AI lifecycle	O			O		O		
AI development lifecycle				O		O		O
AI Ethics Lifecycle								O
AI-enabled system lifecycle	O	O			O	O		
Edge-to-cloud AI lifecycle	O			O	O			
HEAAL	O	O						O
ODD & BC driven AI lifecycle								O
Seven-layer model	O			O	O			O

AI lifecycle model	Model development	Model evaluation	Model testing	Model deployment	Model operation	Model monitoring	Model maintenance	Model documentation
TDSP	O			O				
CRISP-DM	O	O		O				
Microsoft model	O	O		O		O		

Allam's AI lifecycle	O		O	O		O		
End-to-end AI lifecycle	O			O		O	O	
AI application project lifecycle	O		O	O		O		
Refined CRISP-DM	O	O		O		O		O
AI lifecycle in clinical AI	O	O		O				
CDAC AI life cycle		O		O	O	O		
Extended CRISP-DM	O	O		O		O	O	
Isom's AI lifecycle	O	O		O				
Moley's AI lifecycle	O		O	O		O		
Fairness intervention AI lifecycle		O		O				
AI development lifecycle	O	O		O		O		
AI Ethics Lifecycle	O	O			O			
AI-enabled system lifecycle	O		O			O	O	O
Edge-to-cloud AI lifecycle	O	O		O		O	O	
HEAAL				O		O	O	
ODD & BC driven AI lifecycle	O	O		O	O	O		
Seven-layer model	O		O	O				

Table 2. Comparison of stages of AI lifecycle models

Data acquisition aims to obtain data suitable for AI model development. Stakeholders must be aware of data acquisition methods, use privileges, and regulations to make informed decisions. Cybersecurity precautions and import/export policies are necessary when integrating datasets from the open market (Isom, 2022).

Data preparation involves processing and cleaning the acquired data to ensure they are suitable for analysis and model training (Alberti et al., 2024; Haakman et al., 2021). Special attention needs to be taken to avoid biases during the preparation steps, as they can affect the AI model's fairness and accuracy (Agarwal & Agarwal, 2024).

Model design begins with 'a comparison with industry benchmarks and algorithmic baselines' where similar problems have been addressed (De Silva & Alahakoon, 2022, p. 5). It demands a clear definition of the AI algorithm's purpose (Singh, 2022). It then includes, excludes, transforms, or aggregates features for effective and efficient handling of training models (Deck et al., 2024).

Model building involves training suitable AI algorithms that align with the intended application (Alberti et al., 2024). The maturity of AI technologies and the agility of programming languages have contributed to the capability and time efficiency of building several AI models by reimplementing algorithms, which ensures the integrity of AI models (De Silva & Alahakoon, 2022).

Model evaluation indicates models built in the last stage should be evaluated '*using the metrics designed for the type of AI capability and the algorithm*' (De Silva & Alahakoon, 2022), such as the

benchmarks set in the model design stage. When the required capability is not achieved, new algorithms need to be developed based on novel research.

Model deployment involves integrating the AI model into a scenario for operational use (Singh, 2022), setting up necessary interfaces, and ensuring scalability, reliability, and security (Alberti et al., 2024). The best practice to reduce risk is to use safe deployment techniques, including A/B tests and canary releases (Arnold et al., 2020).

Model monitoring is crucial to ensure AI models' accuracy, explainability, and unbiasedness (Isom, 2022) by tracking key performance metrics. Monitoring tools are essential to understand AI actions and mitigate risks, enhancing user confidence through advanced traceability (Alberti et al., 2024).

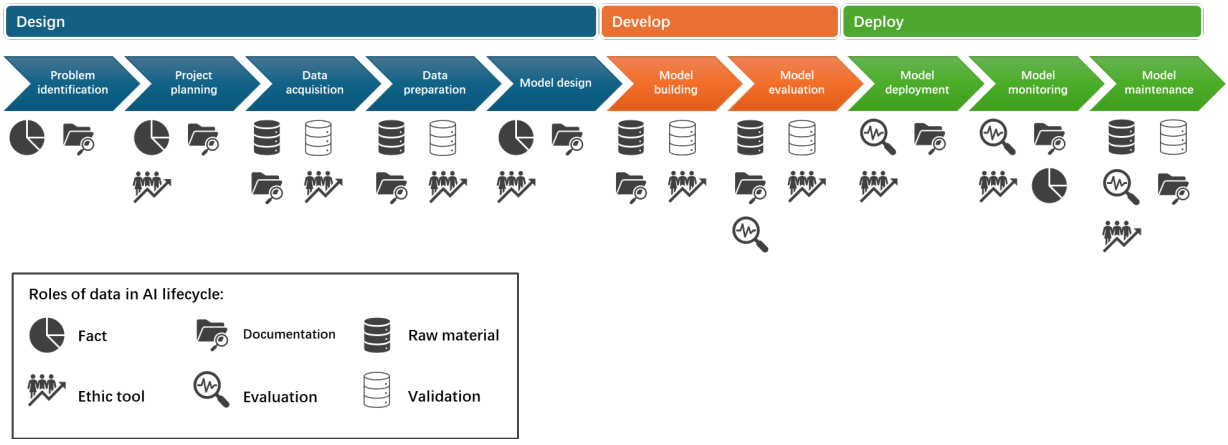


Figure 2. AI lifecycle model with the roles of data

Model maintenance is a stage commonly alternated with the model monitoring stage. It includes handling model drift and addressing issues, which are challenging tasks requiring skills and expertise (Alberti et al., 2024). Many methods can be used to improve AI models, including '*feature engineering, model architecture selection, hyperparameter tuning, and the addition of more training data using techniques such as active learning*' (Arnold et al., 2020, p. 4).

Discussion

To answer RQ3, based on the stages of the AI lifecycle model in the last section, this study identified six roles of data in the whole lifecycle of AI, which are facts, raw materials, validation, evaluation, ethic tools, and documentation. In different stages of the AI lifecycle, data plays different roles (Figure 2). These roles highlight the importance of high-quality, diverse, and well-managed data throughout the AI lifecycle.

Facts mean data describing the environment connected to the AI project. It can help stakeholders understand the problem, thus in problem identification, project planning, model design, and model monitoring stages, stakeholders collect data reflecting the facts to support decisions.

Raw materials mean the data for training AI models, allowing algorithms to learn patterns and make predictions or decisions based on new inputs. It is closely related to *Validation*, which means separating datasets to validate and test the model, ensuring it performs well on unseen data and generalizes effectively. The datasets for training and validating AI models usually are collected and processed together (Singh, 2022), thus datasets play these two roles in data acquisition, data preparation, model building, and model evaluation stages of the AI lifecycle. For model maintenance, sometimes additional datasets are used for the evolvement of AI models (Arnold et al., 2020), as explained in the last section.

Evaluation means data used to reflect the model's performance with various metrics and benchmarks, helping to refine and improve the AI model. AI models are not only evaluated in the development phase to select better-performed ones for the deployment phase (Haakman et al., 2021). They are managed in the deployment, monitoring, and maintenance stages with regular evaluations of their functions (De Silva & Alahakoon, 2022; Isom, 2022).

Ethic tools mean data to be diverse and representative to help in identifying and mitigating ethical issues, including biases (Kim et al., 2024), fairness (Agarwal & Agarwal, 2024), explainability (Singh, 2022), trustworthiness (Solanki et al., 2023), etc. Scholars have been extensively working to address AI ethics issues by inserting bias detection and fairness assessment in all stages of the AI lifecycle (Agarwal & Agarwal, 2024; Calegari et al., 2023), especially for the training datasets and evaluation metrics.

Documentation means the whole lifecycle to be properly documented (Kaas et al., 2024). This includes the evaluation metrics in model evaluation, deployment, monitoring, and maintenance stages, as well as data about the design and the model itself (Kaas et al., 2024). It enables reproducibility and a trail of actions and decisions (Haakman et al., 2021). Sufficient historical data will also help to build a more accurate model with fewer bias issues (Kim et al., 2024). However, data documentation in the AI lifecycle has not yet drawn enough attention. Only two current lifecycles included details of documentation (See Table 2).

Conclusion

This study reviewed 20 AI lifecycle models identified from the current literature. Overlaps of stage contents and confused use of stage names were noticed in these models. A double-layer AI lifecycle model was synthesized from the data-driven perspective, which provides a canonical overview of multiple stages that fit business objectives, AI model development, and various implementing scenarios. We anticipate this AI lifecycle model to address commercial, scholarly, and social interests and contribute to future AI studies. Six roles of data in the AI lifecycle were identified to increase the awareness and emphasis of high-quality, diverse, and well-managed data to support AI model development. This study also shows the documentation function of data in the AI lifecycle has not been fully valued. The key limitation of this study is that the proposed AI lifecycle model with data roles has not been validated with empirical research or practical applications, which calls for future studies to examine the model in real-world settings.

Acknowledgements

This work was supported by the National Natural Science Foundation of China Project No. 72104188 and Project No. 72174153.

About the authors

Di Wang is a Lecturer in the School of Information Resource Management, Renmin University of China, China. She received her double PhDs from Macquarie University in Australia and Wuhan University in China. Her research interests are in big data management, user behaviour, open data utilization, and data services. She can be contacted at di.wang@ruc.edu.cn.

Ruiyang Chen is a PhD student of Information Studies at the School of Information Management at Wuhan University, China. Her research interests include digital trust, open science, and information trust. She can be reached at rychen22@whu.edu.cn.

Chuanni Li is a master student in the School of Information Resource Management, Renmin University of China, China. Her principal research interests lie in the field of human-AI interaction and AI contribute within the library industry. She can be contacted at 1015425325@qq.com.

Shanshan Gu is a master student in the School of Information Resource Management, Renmin University of China, China. Her research focuses on data quality assessment and its interaction with AI. She can be contacted at gushansabel@163.com.

References

Agarwal, A., & Agarwal, H. (2024). A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. AI and Ethics, 4(2), 299-314.

Alberti, E., Alvarez-Napagao, S., Anaya, V., Barroso, M., Barrué, C., Beecks, C., Bergamasco, L., Chala, S. A., Gimenez-Abalos, V., & Graß, A. (2024). AI Lifecycle Zero-Touch Orchestration within the Edge-to-Cloud Continuum for Industry 5.0. Systems, 12(2), 48.

Allam, S. (2020). An exploratory study on why ai lifecycle models need to be revised. Sudhir Allam.(2020). AN EXPLORATORY STUDY ON WHY AI LIFECYCLE MODELS NEED TO BE REVISED. International Journal of Innovations in Engineering Research and Technology, 7(11), 129-134.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP),

Arnold, M., Boston, J., Desmond, M., Duesterwald, E., Elder, B., Murthi, A., Navratil, J., & Reimer, D. (2020). Towards automating the AI operations lifecycle. arXiv preprint arXiv:2003.12808.

Calegari, R., Castañé, G. G., Milano, M., & O'Sullivan, B. (2023). Assessing and enforcing fairness in the AI lifecycle.

Centers of Excellence. (2024). Understanding and managing the AI lifecycle. Retrieved 2025.1.3 from https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/

Cooper, H. (2015). Research synthesis and meta-analysis: A step-by-step approach (Vol. 2). Sage publications.

Corbin, J., & Strauss, A. (2008). Basics of qualitative research: Techniques and Procedures for Developing Grounded Theory (3rd edition). Sage Publications.

Davis, J., Mengersen, K., Bennett, S., & Mazerolle, L. (2014). Viewing systematic reviews and meta-analysis in social research through different lenses. SpringerPlus, 3(1), 511. https://doi.org/10.1186/2193-1801-3-511

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. Patterns, 3(6).

Deck, L., Schoemäcker, A., Speith, T., Schöffer, J., Kästner, L., & Kühl, N. (2024). Mapping the Potential of Explainable Artificial Intelligence (XAI) for Fairness Along the AI Lifecycle. arXiv preprint arXiv:2404.18736.

Douglas, D. (2003). Grounded theories of management: a methodological review. Management Research News, 26(5), 44-52. https://doi.org/10.1108/01409170310783466

El-Haber, N., Burnett, D., Halford, A., Stamp, K., De Silva, D., Manic, M., & Jennings, A. (2024). A Lifecycle Approach for Artificial Intelligence Ethics in Energy Systems. Energies, 17(14), 3572.

Endsley, M. R. (2023). Supporting Human-AI Teams: Transparency, explainability, and situation awareness. Computers in human behavior, 140, 107574.

Ericson, G., Rohm, W. A., Martens, J., Sharkey, K., Casey, C., Harvey, B., & Schonning, N. (2017). Team data science process documentation. Retrieved April, 11, 2019.

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: a cross-disciplinary perspective. IEEE Technology and Society Magazine, 40(2), 72-80.

Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. Daedalus, 151(2), 218-231.

Georgieva, I., Lazo, C., Timan, T., & van Veenstra, A. F. (2022). From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI and Ethics, 2(4), 697-711.

Haakman, M., Cruz, L., Huijgens, H., & Van Deursen, A. (2021). AI lifecycle models need to be revised: An exploratory study in Fintech. Empirical Software Engineering, 26(5), 95.

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. The journal of strategic Information Systems, 32(3), 101772. https://doi.org/https://doi.org/10.1016/j.jsis.2023.101772

Isom, P. (2022). Earning citizen confidence through a comprehensive approach to responsible and trustworthy AI stewardship and governance. Journal of AI, Robotics & Workplace Automation, 1(3), 278-284.

Kaas, M. H., Burr, C., Porter, Z., Ozturk, B., Ryan, P., Katell, M., Polo, N., Westerling, K., & Habli, I. (2024). Fair by design: A sociotechnical approach to justifying the fairness of AI-enabled systems across the lifecycle. arXiv preprint arXiv:2406.09029.

Kim, J. Y., Hasan, A., Kellogg, K. C., Ratliff, W., Murray, S. G., Suresh, H., Valladares, A., Shaw, K., Tobey, D., & Vidal, D. E. (2024). Development and preliminary testing of Health Equity Across the AI Lifecycle (HEAAL): A framework for healthcare delivery organizations to mitigate the risk of AI solutions worsening health inequities. PLOS Digital Health, 3(5), e0000390.

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.-I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J., Wieland, L. S., & Group, o. b. o. t. C. I. R. M. (2019). Searching for and selecting studies. In Cochrane Handbook for Systematic Reviews of Interventions (pp. 67-107). https://doi.org/https://doi.org/10.1002/9781119536604.ch4

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants. Business & Information Systems Engineering, 61(4), 535-544. https://doi.org/10.1007/s12599-019-00600-8

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Flach, P., Hernández-Orallo, J., Kull, M., Lachiche, N., & Ramírez-Quintana, M. J. (2017). CASP-DM: context aware standard process for data mining. arXiv preprint arXiv:1709.09003.

Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: toward human-AI hybrids. MIS quarterly, 43(1), iii-ix.

Rauh, L., Gärtner, S., Brandt, D., Oberle, M., Stock, D., & Bauernhansl, T. (2022). Towards ai lifecycle management in manufacturing using the asset administration shell (aas). Procedia CIRP, 107, 576-581.

Rieder, T. N., Hutler, B., & Mathews, D. J. (2020). Artificial intelligence in service of human needs: pragmatic first steps toward an ethics for semi-autonomous agents. AJOB neuroscience, 11(2), 120-127.

Rzepka, C., & Berger, B. (2018). User interaction with AI-enabled systems: A systematic review of IS research.

Samoili, S., Cobo, M. L., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence.

Singh, J. P. (2022). Human-Centered AI (HCAI) Paradigms in Clinical Artificial Intelligence: An Analytical Discourse on Implementation Across AI Lifecycle Stages. Emerging Trends in Machine Intelligence and Big Data, 14(4), 17-32.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333-339. https://doi.org/https://doi.org/10.1016/j.jbusres.2019.07.039

Solanki, P., Grundy, J., & Hussain, W. (2023). Operationalising ethics in artificial intelligence for healthcare: A framework for AI developers. AI and Ethics, 3(1), 223-240.

Stettinger, G., Weissensteiner, P., & Khastgir, S. (2024). Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. IEEE Access.

Xia, B., Lu, Q., Zhu, L., & Xing, Z. (2024). An AI System Evaluation Framework for Advancing AI Safety: Terminology, Taxonomy, Lifecycle Mapping. Proceedings of the 1st ACM International Conference on AI-Powered Software,

Zancul, E., Rodrigues, V. P., & Rozenfeld, H. (2019). Insights on software selection based on reference models for business processes: an example in product lifecycle management. International Journal of Product Lifecycle Management, 12(2), 131-148. https://doi.org/10.1504/ijplm.2019.107006

Zhou, J., & Chen, F. (2021). Towards humanity-in-the-loop in AI lifecycle. In Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership (pp. 3-13). Springer.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Gallery walk as research method in information science

Rebecca Frank and Stephanie Krueger
DOI: https://doi.org/10.47989/ir30iConf47554

Abstract

Introduction. This paper argues that the gallery walks, a pedagogical technique traditionally used in educational settings, is an effective qualitative data collection strategy. Its interactive format stimulates discussion and ensures active participation, making it suitable for qualitative research.

Background. Common qualitative methods like interviews, focus groups, and observations have limitations. The gallery walk technique leverages movement and interaction to deepen understanding, making it an effective tool for comprehensive and inclusive educational engagement.

Gallery walk as research method. The gallery walk enhances data collection by balancing individual and group insights, encouraging participants to use their expertise and engage in meaningful discussions. This approach captures detailed information from each participant.

Case study. We implemented the gallery walk in a study with 14 experts in satellite image analysis. Over two days, participants engaged with six thematic stations, discussing and annotating posters. The discussions were audio-recorded and transcribed, providing a rich dataset of individual and collective insights.

Discussion and Conclusion. Our findings demonstrate the gallery walk's utility as a qualitative research method. Its structured yet flexible format enhances participant engagement and data richness. The gallery walk is particularly effective for studies involving expert participants, offering a comprehensive understanding of research topics.

Introduction

In this short paper we argue that the gallery walk, a pedagogical technique developed and used in educational settings, is a highly effective qualitative data collection strategy. The gallery walk is an activity in which small groups of people rotate through a physical space, interacting with a series of artifacts, before coming together as a whole to discuss and debrief (e.g., Francek, 2006). This interactivity not only stimulates discussion but also ensures active participation from all individuals. Previous scholarship on the gallery walk highlights its efficacy in promoting dialogue and engagement in the classroom, qualities that are equally beneficial in qualitative research settings (e.g., Hakim et al., 2019).

Traditional research methods each have distinct tradeoffs: interviews provide depth but are time-consuming, focus groups enable efficient group discussion but can limit individual voices, and observations capture valuable context but lack direct participant input. We sought a method balancing these approaches - one that could efficiently gather rich data from a large group while enabling both individual contributions and observation of expert interactions. Our chosen approach combined the key advantages of interviews (detailed individual perspectives), focus groups (efficient large-group data collection), and observational methods (insight into expert dynamics).

In our research, we have adapted the gallery walk for a qualitative research context and implemented it in a study focused on understanding how experts in satellite image analysis make determinations about the trustworthiness of data, the risks associated with data use, and the broader social and political contexts in which they conduct their work. The use of the gallery walk technique allowed us to gather nuanced insights from this group, both in the form of audio recordings of their discussions as well as the annotations they left on each artifact.

Our findings demonstrate that the gallery walk method facilitates in-depth data collection and enriches participant engagement, making it a valuable addition to qualitative research methodologies. By presenting the successful implementation of the gallery walk in our research project, this paper aims to contribute to the ongoing dialogue about innovative qualitative data collection methods. We believe that the gallery walk's structured, yet flexible approach holds significant potential for various research contexts, promoting richer and more inclusive data collection processes.

Background

Qualitative data collection: interview, focus group, observation

Qualitative research encompasses a wide range of data collection and analysis techniques. Qualitative research methods literature frequently emphasizes that it is difficult to settle on a single definition of qualitative research (e.g., Rubin, 2021; Small and Calarco, 2022). Common qualitative data collection techniques used in human subject research include interview, focus group, and observation, (e.g., Creswell, 2009; Hennink et al., 2019; Rubin, 2021). Qualitative data collection can also include the acquisition of text-based artifacts such as documents or web-based content (Rubin, 2021). Qualitative data collection methods are discussed in ways that emphasize examining people's experiences and identifying '*issues from the perspective of your study participants [to] understand the meanings and interpretations that they give to behavior, events or objects*' (Hennink et al., 2019, p. 10).

Discussion- and/or observation-based qualitative data collection techniques can be highly regimented, as with interview procedures that rely on a scripted set of questions from which interviewers cannot deviate (e.g., Tourangeau et al., 2000), or open-ended, as with observations in which researchers capture information about how participants behave in a natural setting (e.g., Creswell, 2009). Between those two extremes are a range of semi-structured interventions in

which researchers have an interview or observation protocol to guide their interactions with participants, but can go off-script as needed to ensure that necessary and/or relevant topics are covered (e.g., Karatsareas, 2022).

These methods for collecting qualitative data can focus on one participant at a time, as with a one-on-one interview, or can allow researchers to bring groups of people together for a single event, such as a focus group. Research can benefit from the singular focus of an interview because researchers are able to obtain high-quality and in-depth information from a focused encounter, while focus groups allow researchers to benefit from a collective experience in which participants are able to interact with one another in addition to the researchers (Hennink et al., 2019). For interviews, the investment of time required to collect data from a large number of participants can be prohibitive, while focus groups can be efficient in terms of time but the amount of information collected from any given participant may be small.

Our research team was interested in finding a data collection method that would allow us to speak with a large group of people in a limited timeframe, while still retaining the ability to collect in-depth information from each. We also sought a data collection method that would allow us to capture interactions between experts as they discussed topics relevant to their expertise. This led us to explore an educational technique: the gallery walk.

Gallery walk

The gallery walk is most often used in educational settings. It is an active learning technique in which participants rotate through a set number of stations, discussing and/or interacting with artifacts at each.

Scholarship about the gallery walk as an instructional method includes guidance about how to implement a gallery walk in an educational setting (e.g., Francek, 2006), descriptions of particular implementations (e.g., Karlsson, 2020; Rodenbaugh, 2015), research about the effectiveness of the gallery walk as an instructional method (e.g., Chin et al., 2015; Hakim et al., 2019), and scholarship about the use of the gallery walk in professional development and teacher education (e.g., McCafferty and Beaudry, 2017; Ramsaroop and Petersen, 2020). An early publication about implementing a gallery walk provides step-by-step instructions for facilitating this type of activity (Francek, 2006). Later publications describe modifications and/or updates to this basic model (Chin et al., 2015; Karlsson, 2020; Rodenbaugh, 2015). Publications about implementing a gallery walk in K-12 education tend to focus on providing instructions (Francek, 2006), while scholarship about implementing a gallery walk in undergraduate and graduate education emphasize evaluating the effectiveness of the instructional method (Chin et al., 2015; Hakim et al., 2019; Karlsson, 2020).

Scholars and educators who have written about their experiences implementing the gallery walk method emphasize that it is particularly effective at encouraging and facilitating discussion between participants (e.g., Chin et al., 2015; Francek, 2006; Hakim et al., 2019; Karlsson, 2020; McCafferty and Beaudry, 2017; Ramsaroop and Petersen, 2020; Rodenbaugh, 2015). Many characterizations frame the gallery walk in opposition to traditional stand-and-deliver or sage-on-a-stage teaching methods, in which students are expected to listen to an instructor lecture about a topic (Chin et al., 2015; Francek, 2006; Hakim et al., 2019; McCafferty and Beaudry, 2017; Ramsaroop and Petersen, 2020; Rodenbaugh, 2015). Others describe it as a complement to traditional teaching methods rather than an alternative (Karlsson, 2020).

Several scholars have noted that participants in a gallery walk bring their own experiences, expertise, and knowledge to the activity (Chin et al., 2015; Francek, 2006; Rodenbaugh, 2015). Others have noted that the gallery walk is a particularly effective setting for exploring open-ended questions (Francek, 2006), and stress the importance of facilitators asking the right kinds of questions and/or presenting information in useful ways in order to ensure a fruitful experience

for participants (McCafferty and Beaudry, 2017). Another key feature of the gallery walk is movement. When implemented in an in-person setting, participants must literally walk around the space as they move from station to station. This is frequently described as a benefit of the activity: '*a gallery walk is a discussion technique that gets learners out of their seats and invites them to become active participants in the learning process*' (McCafferty and Beaudry, 2017, p. 49).

Some have gone so far as to argue that a key benefit of the gallery walk is that it allows participants to show off their knowledge (Ramsaroop and Petersen, 2020), and to practice their speaking skills (Hakim et al., 2019). Indeed, Karlsson explained that

> *unlike during a traditional literature seminar when perhaps only a few students are active and the teacher is the one primarily giving feedback, the gallery walk format ensures that all students actively contribute and give each other feedback that is both situated and interactive* (2020, p. 91).

For a gallery walk to work effectively, participants are not only provided with the opportunity to participate but in fact must actively participate.

Gallery walk as research method

We argue here that the features that make the gallery walk an effective instructional method also make it a promising method for qualitative research. The approach excels through its emphasis on participant preparation and structured interaction, enabling deep engagement from expert participants. Unlike traditional focus groups where some voices may dominate, the gallery walk's format ensures comprehensive participation. This creates an environment where researchers can efficiently gather rich, nuanced perspectives from large groups while maintaining the quality of individual contributions.

In conclusion, the same characteristics that make the gallery walk an effective teaching strategy—its interactivity, focus on preparation, and inclusive nature—also make it a valuable method for qualitative research. By creating an environment where participants actively contribute throughout the data collection process, the gallery walk enables researchers to capture a wide array of nuanced insights from a large group of people in a relatively short span of time.

Case study: gallery walk in information science research

We employed a gallery walk activity as a data collection event for citizen-based monitoring for peace and security in the era of synthetic media and deepfakes project in July 2024. In this implementation, we organized a two-day, in-person workshop for a group of 14 experts. The goal of this event was to examine how people with expertise in satellite image analysis make decisions about trust and risk in the course of their work with satellite data.

The gallery walk format suited our data collection needs given the event parameters and group size. It enabled participants to discuss in smaller groups rather than one large assembly, allowing our research team to gather more diverse perspectives - including from those who might have been hesitant to speak in the larger group setting.

The gallery walk and full group discussions were audio recorded and transcribed for analysis. This research was reviewed and approved by the Institutional Review Board at the first author's institution.

Participants

Gallery walk participants were selected based on their expertise and experience with satellite imagery analysis. The participants primarily had experience in peace and security (e.g., using satellite data for nonproliferation treaty verification, or to monitor the movement of refugees), or in environmental advocacy (e.g., using satellite data to monitor large-scale agricultural trends).

Participants came from a variety of professional affiliations, including academia, journalism, nonprofit/advocacy groups, government research laboratories, and satellite providers. Participants came from North America and Europe.

Gallery walk methodology

Our gallery walk implementation included four phases, with breaks throughout the 2-day event at suitable intervals: (1) introductory discussion, (2) gallery walk rotations, (3) full-group discussion, and (4) debrief. Each phase will be described in brief below.

Introductory discussion

The data collection event began with a whole group discussion. While some participants knew one another prior to this event, a round of introductions and a brief discussion allowed all members of the group to share information about their background, expertise, and interest in satellite imagery. The research team introduced the project, and explained the logistics of the gallery walk.

Gallery walk rotations

Our gallery walk featured six stations for participants to visit during the rotations. Participants were randomly assigned to small groups, and each group started at a different station. They were asked to discuss the poster at each station, which featured a topic, image(s), and brief text prompt. They were given pencils and sticky notes in different colors to annotate the posters at each station. Each group was asked to use the same color sticky notes throughout the gallery walk so that their comments could be reliably tracked by our team. Audio recording devices were placed at each station to capture participant discussions.

The posters at each station were designed by our research team to focus participants' discussion and annotations on current and outstanding issues in satellite image analysis: (1) bearing witness from afar, (2) when there is no ground truth, (3) censorship and deception, (4) synthetic and doctored images, (5) surveillance and privacy, and (6) seeing patterns of life. See Figure 1 below. Each poster was placed at evenly spaced intervals around the perimeter of a large event space.

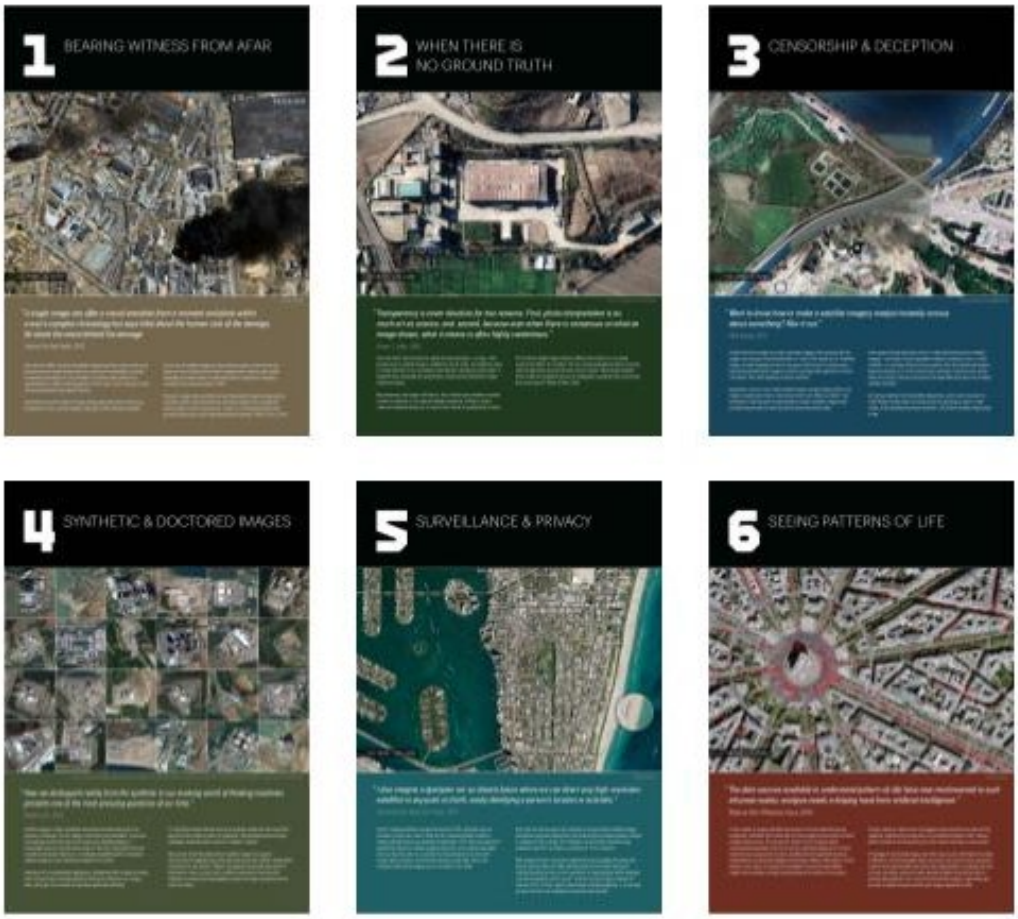


Figure 1. Images of the posters from each gallery walk station

Participants were given approximately 20 minutes at each station and rotated through all six stations in succession. In each rotation, participants were asked to discuss the poster, make annotations using their sticky notes, and were encouraged to engage with both the poster and the annotations left by the other groups. The research team answered logistical questions about the gallery walk procedures as needed, but did not provide information or clarification about the contents of the posters during the gallery walk.

Full-group discussion

Once the group went through six rotations, and every participant had visited every station, we brought everyone together for a full-group discussion. We presented each poster and asked the group that had visited that station last to summarize the topic, their discussion, and the annotations left by all groups at that station before moving into a broader discussion of the topic. The guided full-group discussion addressed a range of topics, and we allowed the participants to direct the conversation based on their insights, interests, questions, and opinions.

Debrief

After each of the six posters had been discussed at length, we held a debrief session for the participants. In the debrief, we shared more information about our research project, including a discussion about how the workshop and gallery walk fit into our work. We gave participants time to ask us questions and to discuss anything that they wanted to share but had not yet been able to discuss.

Participants asked the research team a variety of questions during the debrief. Some were focused on the broad goals of our research project, including questions about our motivation and intended outcomes. Others focused on the gallery walk activity, including questions about the design process for the posters that were placed at each station and the implementation of the gallery walk as a research method. A third category of questions focused on project outputs, including publications, presentations, and data.

We found that our answers to participant questions during the debrief led to additional insights. For example, questions about the design of the gallery walk posters led participants to share additional thoughts about some of the topics and prompts that had not yet been mentioned in either the gallery walk rotations or the full-group discussion.

Discussion and conclusion

The outcomes of our gallery walk demonstrate its potential as a qualitative data collection method. The gallery walk is a structured, interactive technique that allows researchers to collect robust data from a group of participants. It is particularly effective when working with participants who have been selected for their expertise in an area that is relevant to the research. In our case, this meant that we were able to bring together a group of experts in satellite image analysis to lead a guided discussion about issues in satellite image analysis for citizen-based monitoring and societal verification.

By creating an environment for data collection in which participants were asked to actively engage with a set of artifacts and allowing them to both discuss and annotate those artifacts, we were able to create a dynamic and inclusive data collection process in which all participants' voices were heard.

One of the primary advantages that we observed in our implementation of the gallery walk was the balance between individual and group data gathering. We captured every participant's view about each of our six artifacts, as well as the whole-group discussion in which participants were able to engage with one another and the research team. This led to the creation of a dataset that captured both individual insights and collective deliberations. This dual focus has given us the ability to examine how experts in satellite image analysis understand and navigate issues of trust and risk associated with their work.

In conclusion, our study underscores the utility of the gallery walk as a data collection method in qualitative research, particularly for studies involving expert participants. Its emphasis on preparation, interactivity, and inclusivity aligns well with the goals of qualitative research to explore complex phenomena from multiple viewpoints. By fostering an environment where participants are both contributors and learners, the gallery walk method not only gathers valuable data but also enhances participant engagement and investment in the research process. Future research should continue to explore and refine this method, potentially adapting it for use in various disciplinary and interdisciplinary contexts, to further validate its effectiveness and versatility.

Acknowledgements

We would like to thank the participants who generously shared their time and expertise with us. We would also like to thank the following members of our research team, who have made valuable contributions to this research: Dr. Alex Glaser, Princeton University, and Dr. Felix Biessmann and Tuong Vy Nguyen, Berliner Hochschule für Technik. The Einstein Center Digital Future in Berlin, Germany provided space and logistical support for our workshop. This research was funded by a grant from the Deutsche Stiftung Friedensforschung for the '*citizen-based monitoring for peace and security in the era of synthetic media and deepfakes*' project (2023-2025).

About the authors

Dr. Rebecca Frank is an Assistant Professor at the University of Michigan school of information. She is also affiliated with the Einstein Center Digital Future and the Inter-university Consortium for Political and Social Research (ICPSR. She can be contacted at frankrd@umich.edu.

Dr. Stephanie Krueger is a scholar and researcher affiliated with the Berliner Hochschule für Technik, the Czech National Library of Technology, and Czech Technical University in Prague. She can be reached at stephanie.krueger@gmail.com.

References

Chin, C.K., Khor, K.H., Teh, T.K., 2015. Is Gallery Walk an Effective Teaching and Learning Strategy for Biology?, in: Gnanamalar Sarojini Daniel, E. (Ed.), Biology Education and Research in a Changing Planet. Springer Singapore, Singapore, pp. 55–59. https://doi.org/10.1007/978-981-287-524-2_6

Creswell, J.W., 2009. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage, Los Angeles.

Francek, M., 2006. Promoting Discussion in the Science Classroom Using Gallery Walks. Journal of College Science Teaching 36, 27–31.

Hakim, M.A.R., Anggraini, N., Saputra, A., 2019. Gallery Walk Technique to Improve Students' Speaking Skill. scr. j. j. linguist. engl. teach. 4, 26–37. https://doi.org/10.24903/sj.v4i1.251

Hennink, M., Hutter, I., Bailey, A., 2019. Qualitative Research Methods, 2nd ed. SAGE Publications Ltd, Thousand Oaks.

Karatsareas, P., 2022. Semi-Structured Interviews, in: Kircher, R., Zipp, L. (Eds.), Research Methods in Language Attitudes. Cambridge University Press, pp. 99–113. https://doi.org/10.1017/9781108867788.010

Karlsson, R., 2020. Gallery Walk Seminar: Visualizing the Future of Political Ideologies. Journal of Political Science Education 16, 91–100. https://doi.org/10.1080/15512169.2018.1549495

McCafferty, A., Beaudry, J., 2017. The gallery walk: Educators step up to build assessment literacy. The Learning Professional 38, 48–53.

Ramsaroop, S., Petersen, N., 2020. Building Professional Competencies Through a Service Learning 'Gallery Walk' in Primary School Teacher Education. JUTLP 17. https://doi.org/10.53761/1.17.4.3

Rodenbaugh, D.W., 2015. Maximize a team-based learning gallery walk experience: herding cats is easier than you think. Advances in Physiology Education 39, 411–413. https://doi.org/10.1152/advan.00012.2015

Rubin, A.T., 2021. Rocking Qualitative Social Science: An Irreverent Guide to Rigorous Research. Stanford University press, Stanford (Calif.).

Small, M.L., Calarco, J.M., 2022. Qualitative Literacy: A Guide to Evaluating Ethnographic and Interview Research. University of California Press, Oakland (Calif.).

Tourangeau, R., Rips, L.J., Rasinski, K.A., 2000. The Psychology of Survey Response. Cambridge University Press, Cambridge, U.K.; New York.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Discourses of fear around AI and their implications for library and information science

Sarah Appedu and Yigang Qin
DOI: https://doi.org/10.47989/ir30iConf47548

Abstract

Introduction. Since its inception, the seemingly unlimited potential of artificial intelligence (AI) to alter human existence has evoked feelings of fear and amazement. Today, all sectors of industry, academia, and society are anticipating the potential changes new AI technologies are forecasted to bring and mitigate their harms. In this climate, there is a clear need to centre the complex interactions between discourse, power, and individual/institutional actors within sociotechnical systems and their material consequences.

Theoretical framing. While scholars have previously made connections between discourses of fear and library and information science (LIS), there has not yet been an attempt to understand how discourses of fear may currently be shaping the field's response to AI. In this paper, we argue that focusing our critical gaze on the discourses of fear shaping the material interactions between LIS, technological artifacts, industry, and society better positions us to intervene in the predicted trajectory of AI innovation.

Conclusion. We posit that cultivating discourses of refusal – which are committed to the belief that more just worlds must be possible – requires both individual and collective consideration of how fear has and continues to shape our own responses to new technologies.

Introduction

Since its inception in the minds of scientists, industrialists, and sci-fi writers, the seemingly unlimited potential of artificial intelligence (AI) to alter human existence has evoked feelings of fear and amazement (Cave & Dihal, 2019). Today, all sectors of industry, academia, and society are anticipating the potential changes new AI technologies are forecasted to bring and mitigate their predicted and current harms. In this climate, there is a clear need to move past sensationalized narratives around the utopian and dystopian possibilities of AI '*monsters*' and centre the complex interactions between discourse, power, and individual/institutional actors within sociotechnical systems and their material consequences for human, environmental, and technological agency (Barad, 2007; Haraway, 2002). Fear plays an important role in these interactions, where it may be simultaneously triggered by personal experiences with new technologies that push the boundaries of taken-for-granted binaries like human/machine and by the discourses of fear entangled with these artifacts. While fear has previously been discussed within the context of new technologies (i.e., '*technophobia*' (Brosnan, 1998).), it is imperative that these fears are situated within the sociotechnical arrangements that incite these feelings and the actors who benefit from them. In our view, dystopian and utopian narratives around AI can be summarized as discourses of fear (Altheide & Michalowski, 1999), which aim to shape the social, political, and economic arrangements that support the continued dominance of '*technological elites*' (Noble & Roberts, 2019) and the long-standing systems of power that sustain them (Barbrook & Cameron, 1996; Benjamin, 2022).

While scholars have previously made connections between discourses of fear and library and information science (LIS) (Radford & Radford, 1997, 2001), there has not yet been an attempt to understand how discourses of fear may currently be shaping the field's response to AI. Indeed, in its growing emphasis on data science-based research and curricula (Wang, 2018), LIS is inherently implicated not only in present discourse around AI, where students and faculty assimilate into and resist cultures of innovation (Hilbert, 2018) but its history, serving as part of the epistemological and ontological foundation on which AI was designed (Maruyama, 2021). Therefore, our framing of concerns and aspirations around AI as discourses of fear attunes us toward classification and control as central mechanisms of this discourse that reinforce the images of monstrousness on which these discourses rely.

In this paper, we argue that focusing our critical gaze on the discourses of fear shaping the material interactions between LIS, technological artifacts, industry, and society (among others, including the military and non-human environment) better positions us to intervene in the predicted trajectory of AI innovation by moving past sensationalized narratives and centring the agency of individuals and institutions as responsible for technological change and its consequences in every regard. The literature we utilize is not meant to be exhaustive and serves to ground our theorizing of discourses of fear around AI and LIS in the scholarship we have encountered thus far as researchers and practitioners in computer science, philosophy, science and technology studies, and LIS. We also do not mean to dismiss fear as a natural and potentially reasonable response to new technologies (McClure, 2018), a fear that is evoked not only by fear-based discourses but from the lived experiences of harm that result from them and their predecessors (Bender et al., 2021). Instead, we hope to draw more acute attention to how LIS scholars and professionals shape and are shaped by discourses of fear and inspire our communities to take agency over our role in these arrangements and actively participate in discourses of resistance and refusal.

Expanding prior conceptualizations of discourse(s) of fear

Fear has historically been wielded by those in power as an influential motivator for conforming to dominant ideologies (Altheide, 2017; Ericson et al., 1993). The term '*discourse of fear*' originates in the communication to articulate how the media intentionally perpetuates fear-based narratives to undermine readers' sense of agency and reinforce systems of social control that claim to offer

protection from the unknown. Altheide and Michalowski (1999) conceptualize a discourse of fear as '*the pervasive communication, symbolic awareness, and expectation that danger and risk are a central feature of the environment*' (p. 476). Fear has already been discussed as a material-discursive phenomenon in the context of new technologies (Gleason, 2014; Goertzel, 2015; Kothare-Arora, 2020; McClure, 2018). Jassanoff & Kim (2009) theorize the concept of '*sociotechnical imaginaries*' to analyse how technological development is intertwined with national interests, including the leveraging, and containing of what we would call discourses of fear to influence public perception of potentially dangerous technologies, like nuclear weapons. Grenham (2020) discusses the public's '*fear and fascination*' (p. 8) after computers entered the mainstream, sparked in part by science fiction stories and in part by their perception that computers replicated abilities previously reserved for humans. The technology industry worked with the media to counteract these fears and convince people to bring computers into their homes and workplaces by portraying them as subservient to the human operator. Discourses of fear have also been invoked to theorize the material implications of libraries and knowledge organization systems, though this is less common. Radford & Radford (2001) conceptualize discourse of fear from a Foucauldian perspective to examine portrayals of librarians and libraries as enforcers of social control. Finally, Andersen and Skoouvig (2006) use a similar framework to assess the discursive power of knowledge organization systems and better account for how they shape political and social realities.

These perspectives represent valuable contributions to our theorization of discourses of fear surrounding AI. However, they do not always account for the multi-layered and entangled nature of these discourses (Barad, 2007), ignoring their co-constitutive nature across academic and professional sectors, nor do they always explicitly acknowledge the actors that manipulate these discourses to evade accountability and profit off their consequences. Therefore, we expand prior conceptualizations to theorize this phenomena as multiple, dynamic discourses of fear, which simultaneously attunes us toward the intersecting dimensions of technological artifacts, the rhetoric through which they are presented, the actors that create and adopt them, and the social, political, and economic factors that shape their development. Furthermore, theorizing discourses of fear requires acute attention to and accountability for their harms and consequences, both tangible and predicted. Therefore, to contribute to critical conversations meaningfully and ethically around AI and its impacts, we must understand how discourses of fear may be resisted and refused within our own research activities in LIS, particularly regarding the quickly growing conversations about the influence of new AI technologies and the possible worlds they might create or prevent.

Next, we describe how narratives around AI function as discourses of fear. While a full investigation of these discourses within LIS is beyond the scope of this paper, we suggest implications of discourses of fear around AI on the field and directions for future empirical research.

Theorising discourses of fear around AI

A sentient, all-knowing, and often hostile AI entity has been the subject of sci-fi novels and films for decades and is increasingly becoming a common way of discussing the AI technologies now being released for public use (Compagna & Steinhart, 2020; Dishon, 2024). For many, the AI '*monster*' is turning from a product of pure imagination into a very real and concerning possibility. For example, the New York Times published a story about a reporter's '*scary*' interaction with Bing's then newly released AI chatbot. While their conversation started out harmlessly enough, the author ultimately felt disturbed by the encounter after the bot told him it loved him and tried to convince him to leave his wife (Roose, 2023).

However, these fears do not just stem from sci-fi stories and personal encounters with unfamiliar and aggressive AI technologies. Industry leaders are responsible for participating in this monstrous rhetoric, like Elon Musk in his contradictory moves of first warning against summoning an AI '*demon*' (McFarland, 2016) to investing billions of dollars into its creation (NBC News, 2024). These

leaders reinforce feelings of fear through dystopic images of the world if AI were to become sentient, including the risk of human extinction (Bostrom, 2002; Ord, 2020), while paradoxically promoting dystopian views of the world if technologists let these risks stop them from advancing AI, including economic collapse, social unrest, and climate disaster – as if most people globally are not already experiencing these near-apocalyptic conditions (Ali, 2019). Concentrating the public's fear on the AI '*monster*' is not only useful for generating support for their continued control over this trajectory as the supposedly only ones with the knowledge and money to save humanity from the '*existential risks*' associated with AI development. Discourses of fear also conveniently distract from the material harms being perpetuated by the leaders of this industry and the social, political, and economic systems of inequality that sustain their efforts.

Critical scholars have revealed how fear is deliberately perpetuated by discourses that catalyse AI technologies' development and implementation (Cave et al., 2019; Cave & Dihal, 2019; Nguyen & Hekman, 2022; Wajcman, 2017). Several have recognized the legacies of racism, capitalism, and colonialism that have shaped AI innovation and their leveraging of discourses of containment, classification, and control to justify their pursuits (Ali, 2019; Benjamin, 2019; Eubanks, 2018; Tacheva & Ramasubramanian, 2023). In their landmark paper, Torres & Gebru (2024) '*trace the goal of building AGI to the Anglo-American eugenics movement*' (p. 2). While they do not name fear as an element of the TESCREAL bundle, its apocalyptic predictions, where AI is simultaneously the harbinger of Armageddon and the sole saviour of humanity from all imaginable apocalyptic futures, implicitly evoke fear to convince the public that their lives are in technologists' hands. A supposedly alternative conception of AI that may appear to remedy this discourse is techno-optimism, where technological advancement is seen as '*...the glory of human ambition and achievement, the spearhead of progress, and the realization of our potential*' (Andreessen, 2023). However, the logic underlying these sentiments maintains capitalism, colonialism, and heteropatriarchy as the necessary mechanisms for human prosperity and similarly relies on long-standing discourses of fear to justify this claim (Alexander & Rutherford, 2019). These critical perspectives can be understood as revealing both the deep historical roots AI innovation shares with other social and political discourses of fear that serve the ends of containment, classification, and control, as well as the present-day consequences that may be exacerbated if these discourses are not interrogated within our own research practices in LIS.

Implications for library and information science

Given the dire consequences associated with AI-related discourses of fear, we must take a critical look at how LIS as material and discursive fields of scholarship and practice may be implicated in discourses of fear to successfully resist them. As mentioned previously, the fields of LIS are deeply entangled with the technology industry and its '*innovations*' through our professional practices, research epistemologies, curriculum, funding sources, and iSchool-to-industry professional pipeline (Detlefsen, 2007; Hilbert, 2018; Maruyama, 2021). Some prior examples examining discourses of fear within librarianship and knowledge organization exist in LIS literature, yet little attention has been paid to how discourses of fear may be presently shaping and shaped by our professional and personal responses to AI. Notably, Radford & Radford (2001) described libraries as devoted to discursive domination, particularly through practices of classification and surveillance that serve to control patrons' behaviours. They assert that '*it is not the contents of the library that inspire awe; it is the practices that occur within the institution*' (p. 303). In today's climate, we must consider how library's use of and response to AI may echo this history of discursive control and order.

One way to conceive of librarians' role in the AI era is as responsible for preserving the moral order of a democracy threatened by increased mis- and disinformation by teaching their communities digital and information literacy skills, including those needed to critically use AI and navigate its impacts (Cooke, 2018; Jaeger & Taylor, 2021; Ridley & Pawlick-Potts, 2021). While this mission aligns

with librarians' professional values, we should be wary of how discourses of fear may shape our narratives about AI and render patrons ultimately powerless to confront AI without the intervention of the librarian. Ettarh (2018) coined the term '*vocational awe*' to summarize the ways in which librarians historically have held saviour roles in their communities that were defined by the racialized and gendered expectations of upper-class white women preserving the social order of racism and patriarchy. These roles can be reprised in discourses around AI that prioritize the ways in which information professionals may '*save*' society from the harms of AI without interrogating how we may be complicit in them. Instead, librarians should consider how they might resist and refuse AI adoption in their local contexts by recognizing their own agency within complex sociotechnical systems (Appedu & Tacheva, in press).

Beyond librarianship, knowledge organization and information management scholarship and practice must be interrogated for how they may implicitly converse with the same systems of power that inform discourses of fear around AI. While classification may be vital to information organization, access, and retrieval (Kwasnik, 1999), its discursive and material association with fear and control cannot be overlooked (Tacheva & Ramasubramanian, 2023; Torres & Gebru, 2023). Anderson & Skouvig (2006) previously acknowledged that the role of information classification in reinforcing social and political structures has been overlooked among information scientists, thereby allowing them to evade accountability for their role in preserving discourses of fear. They advocate for a '*change in consciousness*' (p. 318), imploring scholars in information management to attend to the social and political order of society as the fundamental knowledge organization system and take ownership for how this informs our ontologies and the technologies that render them material. We agree with Hauser's (2023) portrayal of sociotechnical information systems, or '*systems of record*', '*as a critical site for projects that hold justice to be a central concern within information studies*' (p. 2) and argue that conceptualizing these systems as part of larger discourses of fear allows us to direct our attention towards the interlocking and mutually constitutive nature of information systems, technological designers, knowledge managers, and society. Future research must investigate how else discourses of fear may manifest in and contribute to our work across the information sciences utilizing theoretical frameworks that acknowledge not only the entangled agencies of humans, technologies, and social and natural structures, but the ethical obligations they make visible (Barad, 2007; Haraway, 2002; Ricaurte, 2023). Such work is necessary for not just living in, but actively and ethically shaping our algorithmic, yet always material, world.

Conclusion

This paper contributes to discussions about the implications of AI on LIS research and practice through critically conceptualizing discourses of fear as a material-discursive strategy that aims to enact social control through various sociotechnical systems. Among others, Radford & Radford (2001), have recognized that these discourses reveal a more potent fear held by those that curate them – that the very groups over which technological elites desire control will recognize their capacity to resist. As they argued two decades ago, '*The fear [of those in power] is of the potential to transgress dominant discursive forms and of making all things and all worlds possible*' (p. 307). If we take this to be true, the power of individuals working in collective solidarity to shape technological futures becomes clear. What would it look like to accept Benjamin's (2019) '*invitation to refuse the illusion of inevitability in which technologies of race come wrapped and to 'hotwire' more habitable forms of social organization in the process*' (p. 23)? These questions cause us to reckon with the reality that, while it has certainly led to new ways of building community (Kennedy, 2018), technological advancement has always been used to maintain the power of the elite and disempower those on whose subjugation they rely. Furthermore, they allow us to embody the many other possible responses to new technologies beyond fear, including resistance and refusal. We posit that cultivating discourses of refusal – which are committed to the belief that more just worlds must be possible (Nxumalo, 2021; Simpson, 2007, 2014) – requires both individual and

collective consideration of how fear has and continues to shape our own responses to new technologies.

We have already seen how these discourses manifest in our field through our studies, professional activities, and personal conversations. Future directions for this research include systematic examination of the impact of discourses of fear within LIS and the ways in which information professionals across sectors are already participating in this resistance. This research will allow us to further develop our theorization of discourses of fear around AI and its implications for LIS, as well as contribute to the broader global movement of resistance against oppressive systems within our professional work. Methodologically, this paradigmatic requires a diffractive approach that shifts our gaze from actors and structures and sees technological change as a complex apparatus for which we are all accountable (Barad, 2007). While computers may be incapable of '*thinking*' beyond classification, people working in solidarity have the potential to imagine new ways of orienting ourselves to the world and to one another that do not overlook our deeply entangled futures, seeing technological innovation as inseparable from our quest for liberation from oppressive systems (Gleason, 2014; Haraway, 2002).

Acknowledgements

We would like to thank our colleagues at the Critical AI Research and Education Lab at Syracuse University (Dr. Jasmina Tacheva, Mirakle Wright, and Jeongbae Choi) for their feedback on early drafts of this paper. We also thank the reviewers for their detailed insights.

About the authors

Sarah Appedu is a third-year Doctoral student in Information Science and Technology at the School of Information Studies, Syracuse University. She received her master's in library and information science from the University of Illinois, Urbana-Champaign. Her research interests include feminist and decolonial science and technology studies and critical library, information, & data studies. She can be contacted at spappedu@syr.edu.

Yigang Qin is a second-year Doctoral student in Information Science and Technology at the School of Information Studies, Syracuse University. He earned his B.S. in Computer Science at City University of Hong Kong. His research falls into the intersection of human-computer interaction (HCI) and science and technology studies (STS), specifically on issues of labor, gender, and political economy of technologies. He can be contacted at yqin27@syr.edu.

References

Alexander, S., & Rutherford, J. (2019). A critique of techno-optimism: Efficiency without sufficiency is lost. In *Routledge Handbook of Global Sustainability Governance*. Routledge.

Ali, S. M. (2019). White Crisis' and/as 'Existential Risk,' or the Entangled Apocalypticism of Artificial Intelligence. Zygon®, 54(1), 207–224. https://doi.org/10.1111/zygo.12498

Altheide, D. L. (2017). Creating Fear: News and the Construction of Crisis. Routledge. https://doi.org/10.4324/9780203794494

Altheide, D. L., & Michalowski, R. S. (1999). Fear in the News: A Discourse of Control. The Sociological Quarterly, 40(3), 475–503.

Andersen, J., & Skouvig, L. (2006). Knowledge Organization: A Sociohistorical Analysis and Critique. The Library Quarterly: Information, Community, Policy, 76(3), 300–322. https://doi.org/10.1086/511139

Andreessen, M. (2023, October 16). The Techno-Optimist Manifesto. Andreessen Horowitz. https://a16z.com/the-techno-optimist-manifesto/

Appedu, S. & Tacheva, J. (In press). Transcending Binaries of Agency through Librarians' Discursive Representations of AI. Library Trends 73(3).

Barad, K. (2007). Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. Duke University Press.

Barbrook, R., & Cameron, A. (1996). The Californian ideology. Science as Culture, 6(1), 44–72. https://doi.org/10.1080/09505439609526455

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Polity.

Benjamin, R. (2022). Viral Justice. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691222882/viral-justice

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. Journal of Evolution and Technology, 9. https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c

Brosnan, M. J. (1998). Technophobia: The Psychological Impact of Information Technology. Routledge. https://doi.org/10.4324/9780203436707

Cave, S., Coughlan, K., & Dihal, K. (2019). "Scary Robots": Examining Public Responses to AI. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 331–337. https://doi.org/10.1145/3306618.3314232

Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. Nature Machine Intelligence, 1(2), 74–78. https://doi.org/10.1038/s42256-019-0020-9

Compagna, D., & Steinhart, S. (2020). Monsters, Monstrosities, and the Monstrous in Culture and Society. Vernon Press.

Cooke, N. A. (2018). Fake News and Alternative Facts: Information Literacy in a Post-Truth Era. ALA Editions.

Detlefsen, E. G. (2007). The pipeline problem: Where do we go from here? Journal of the Medical Library Association, 95(2), 115–116. https://doi.org/10.3163/1536-5050.95.2.115

Dishon, G. (2024). From Monsters to Mazes: Sociotechnical Imaginaries of AI Between Frankenstein and Kafka. Postdigital Science and Education. https://doi.org/10.1007/s42438-024-00482-4

Elon Musk's AI start-up now valued at $24 billion after fresh funding. (2024, May 27). NBC News. https://www.nbcnews.com/business/business-news/elon-musks-ai-startup-now-valued-24-billion-fresh-funding-rcna154171

Ericson, R., Baranek, P., & Chan, J. (1993). Negotiating Control: A Study of News Sources. Canadian Journal of Communication, 18(1). https://doi.org/10.22230/cjc.1993v18n1a733

Ettarh, F. (2018). Vocational Awe and Librarianship: The Lies We Tell Ourselves. In the Library with the Lead Pipe. https://www.inthelibrarywiththeleadpipe.org/2018/vocational-awe/

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. First Monday. https://doi.org/10.5210/fm.v29i4.13636

Gleason, S. C. (2014). Don't Fear the Cyborg: Toward Embracing Posthuman and Feminist Cyborg Discourses in Teacher Education and Educational Technology Research. Canadian Journal of Science, Mathematics and Technology Education, 14(2), 120–134. https://doi.org/10.1080/14926156.2014.903320

Goertzel, B. (2015). Superintelligence: Fears, Promises and Potentials: Reflections on Bostrom's Superintelligence, Yudkowsky's from AI to Zombies, and Weaver and Veitas's 'Open-Ended Intelligence.' Journal of Ethics and Emerging Technologies, 25(2), 55–87. https://doi.org/10.55613/jeet.v25i2.48

Grenham, H. (2020). The mechanical monster and discourses of fear and fascination in the early history of the computer. Humanities and Social Sciences Communications, 7(1), 1–11. https://doi.org/10.1057/s41599-020-00650-4

Haraway, D. (2002). 'A Manifesto for Cyborgs: Science, Technology and Socialist Feminism in the 1980s' (1985). In B. Nicol (Ed.), Postmodernism and the Contemporary Novel (pp. 396–420). Edinburgh University Press.

Hauser, E. (2023). Facts in the machine: Systems of record and the performance of sociotechnical truth. Journal of the Association for Information Science and Technology, n/a(n/a). https://doi.org/10.1002/asi.24820

Hilbert, M. (2018). Post-human condition–Epistemic disruption: How Information (Science) lost its body to Data-Based Knowledge. The Future of Education in Information Science, 86–93.

Jaeger, P. T., & Taylor, N. G. (2021). Arsenals of Lifelong Information Literacy: Educating Users to Navigate Political and Current Events Information in World of Ever-Evolving Misinformation. The Library Quarterly, 91(1), 19–31. https://doi.org/10.1086/711632

Jasanoff, S., & Kim, S.-H. (2009). Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea. Minerva: A Review of Science, Learning & Policy, 47(2), 119–146. https://doi.org/10.1007/s11024-009-9124-4

Kothare-Arora, M. (2020). Our Fear of AI: Exploring Its Creators and Creations in Fiction. https://hdl.handle.net/2152/84406

Kwasnik, B. H. (1999). The Role of Classification in Knowledge Representation and Discovery'. Library Trends, 48(1), 22–47.

Maruyama, Y. (2021). Post-Truth AI and Big Data Epistemology: From the Genealogy of Artificial Intelligence to the Nature of Data Science as a New Kind of Science. In A. Abraham, P. Siarry, K. Ma, & A. Kaklauskas (Eds.), Intelligent Systems Design and Applications (pp. 540–549). Springer International Publishing. https://doi.org/10.1007/978-3-030-49342-4_52

McClure, P. K. (2018). 'You're Fired,' Says the Robot: The Rise of Automation in the Workplace, Technophobes, and Fears of Unemployment. Social Science Computer Review, 36(2), 139–156. https://doi.org/10.1177/0894439317698637

McFarland, M. (2016, February 5). Elon Musk: 'With artificial intelligence we are summoning the demon.' Washington Post.

https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/

Nguyen, D., & Hekman, E. (2022). The news framing of artificial intelligence: A critical exploration of how media discourses make sense of automation. AI & SOCIETY. https://doi.org/10.1007/s00146-022-01511-1

Noble, S. U., & Roberts, S. T. (2019). Technological Elites, the Meritocracy, and Postracial Myths in Silicon Valley. In 6 Technological Elites, the Meritocracy, and Postracial Myths in Silicon Valley (pp. 113–130). Duke University Press. https://doi.org/10.1515/9781478003250-007

Nxumalo, F. (2021). Disrupting Anti-Blackness in Early Childhood Qualitative Inquiry: Thinking With Black Refusal and Black Futurity. Qualitative Inquiry, 27(10), 1191–1199. https://doi.org/10.1177/10778004211021810

Ord, T. (2020). The precipice: Existential risk and the future of humanity. Hachette Books.

Radford, G. P., & Radford, M. L. (2001). Libraries, Librarians, and the Discourse of Fear. The Library Quarterly, 71(3), 299–329. https://doi.org/10.1086/603283

Radford, M. L., & Radford, G. P. (1997). Power, Knowledge, and Fear: Feminism, Foucault, and the Stereotype of the Female Librarian. The Library Quarterly, 67(3), 250–266. https://doi.org/10.1086/629951

Ricaurte, P. (2023). AI for/by the majority world: From technologies of dispossession to technologies of radical care. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 3–4. https://doi.org/10.1145/3600211.3607544

Ridley, M., & Pawlick-Potts, D. (2021). Algorithmic Literacy and the Role for Libraries. Information Technology and Libraries, 40(2), Article 2. https://doi.org/10.6017/ital.v40i2.12963

Roose, K. (2023, February 16). A Conversation With Bing's Chatbot Left Me Deeply Unsettled. The New York Times. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

Simpson, A. (2007). On Ethnographic Refusal: Indigeneity, 'Voice' and Colonial Citizenship. Junctures: The Journal for Thematic Dialogue, 9, Article 9. https://junctures.org/junctures/index.php/junctures/article/view/66

Simpson, A. (2014). Mohawk Interruptus: Political Life Across the Borders of Settler States. Duke University Press.

Tacheva, J., & Ramasubramanian, S. (2023). AI Empire: Unraveling the interlocking systems of oppression in generative AI's global order. Big Data & Society, 10(2), 20539517231219241. https://doi.org/10.1177/20539517231219241

Wajcman, J. (2017). Automation: Is it really different this time? The British Journal of Sociology, 68(1), 119–127. https://doi.org/10.1111/1468-4446.12239

Wang, L. (2018). Twinning data science with information science in schools of library and information science. Journal of Documentation, 74(6), 1243–1257. https://doi.org/10.1108/JD-02-2018-0036

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Transforming bridges into smart infrastructure: a data-driven approach to monitoring bats and promoting sustainable coexistence

Jacob Koch, Jess Kropczynski, Janette Perez-Jimenez, and Joseph S. Johnson
DOI: https://doi.org/10.47989/ir30iConf47542

Abstract

Introduction. Smart infrastructure development often prioritizes human needs such as safety and efficiency, overlooking the potential to facilitate coexistence with wildlife. This paper addresses this gap by demonstrating innovative use of internet of things (IoT) technology and data visualization for ecological monitoring.

Method. We repurposed commercially available wildlife trail cameras to collect data on bats roosting in bridges, despite commercial devices not being built for this application. Furthermore, we developed a map- and cloud-based dashboard to integrate this ecological data with existing bridge metadata, transforming raw sensor data into actionable insights for transportation authorities.

Analysis. We explore the ability of the dashboard to meet decision-making needs and offer design implications to further advance wildlife-inclusive smart infrastructure efforts.

Results. Our findings demonstrate the efficacy of using IoT devices for wildlife monitoring and the power of data visualization in informing conservation efforts within smart infrastructure development.

Conclusion. This study showcases the potential of IoT-enabled cameras and a user-friendly dashboard, BatMap, for effective bat monitoring and informed decision-making in infrastructure projects.

Introduction

Smart infrastructure, with its ability to collect and analyze real-time data on various aspects of urban environments, has significantly enhanced the decision-making capabilities of city officials and researchers (Hoult et al., 2009; Rejeb et al., 2022). This interconnectedness, combined with AI, enables the optimization of energy efficiency, informed city development planning, and improved navigation and overall human experience. However, the current focus of smart infrastructure and smart cities is often predominantly human-centric, overlooking the potential to benefit wildlife and promote sustainability (Azevedo Guedes et al., 2018; Mayrand & Clergeau, 2018).

Cities, while built with the intention of supporting humans, also serve as habitats for a variety of plant and wildlife species, contributing to both human well-being and the preservation of natural heritage (Forman, 2014). This realization aligns with the 'more-than-human' perspective in human-computer interaction (HCI), which challenges the traditional anthropocentric design focus and advocates for acknowledging the interconnectedness of humans, technology, and the natural world (Vallgårda & Redström, 2017). Embracing this perspective necessitates understanding how non-human inhabitants interact with and utilize urban spaces.

Such understanding is crucial not only for fostering coexistence but also for avoiding conflicts and seizing opportunities for conservation. For instance, critical infrastructure like bridges often serve as roosting sites for bats (Bektas et al., 2018). This can complicate maintenance projects due to the protected status of many bat species. However, recognizing their presence also highlights the potential for a more-than-human community within our cities (Bektas et al., 2018; Neuman, 2020). Current methods for monitoring bat presence, primarily visually inspecting bridges for bats, have limitations in terms of efficacy (Civjan et al., 2017; Adam & Hayes, 2000; Hendricks et al., 2005).

The challenges facing cities today, including climate change, necessitate innovative solutions and proactive governance. However, the lack of reliable, longitudinal data on urban wildlife populations hinders evidence-based long-term planning. To address this gap, we developed a passive monitoring approach using commercially available trail cameras to transform bridges into smart infrastructure, providing continuous and non-intrusive monitoring and video coverage of wildlife. The collected data is centralized in a cloud-based system, featuring a dashboard interface we call BatMap, which enables visualization, filtering, and facilitates user management.

This paper delves into the necessity for a decision-making tool like ours, employing a case study to illuminate our design process and the reasoning behind its specific features. The resulting design implications reveal how this system can enhance decision-making by furnishing reliable data on bat species, their activity patterns, and suggesting optimal timing for infrastructure maintenance to minimize the impact on wildlife.

The study underscores the potential of IoT-enabled cameras for wildlife monitoring, although it acknowledges the need for further optimization to boost detection rates. It further highlights the effectiveness of BatMap's interface and features in bolstering decision-making processes, particularly through its spatial and attribute-based information presentation. Moreover, the study emphasizes the importance of a 'more-than-human' design perspective in creating technology that benefits both humans and the environment. Finally, the potential of longitudinal data collection to inform conservation efforts and guide the development of wildlife-friendly infrastructure is also brought to light. The research aligns with the iSchools' Climate Action Coalition's focus on promoting interdisciplinary research for climate resilience and sustainable coexistence (iSchools, n.d.).

Specifically, we investigate:

RQ1: to what extent can IoT-enabled cameras and a cloud-based dashboard system effectively detect and monitor bat activity in bridges?

RQ2: how can the map-based dashboard interface support decision-making processes related to infrastructure maintenance and wildlife conservation?

Background

Smart infrastructure: a data-driven foundation for cities

Smart infrastructure, defined as technology capable of sensing its environment and taking automated action based on collected data (Ogie et al., 2017), has become integral to modern urban development. This distinguishes it from intelligent infrastructure, which generates insights requiring human intervention for action (Ogie et al., 2017; Jang et al., 2019; Chun & Ryu, 2019). The confluence of smart and intelligent infrastructure, coupled with the Internet of Things (IoT), allows cities and their infrastructure to become more efficient and responsive to users' needs (Rejeb et al., 2022).

Through a network of sensors deployed across various environments, smart infrastructure systems gather and analyze data to optimize functions like energy efficiency, infrastructure maintenance planning, and even traffic flow management. Technologies like RFID, wireless sensors, and data analytics facilitate real-time monitoring and control of resources such as lighting, public transit, and traffic patterns (Atzori, Iera, & Morabito, 2010; Wilson et al., 2015). Large-scale smart city and infrastructure projects have introduced diverse technologies, opening a plethora of innovative solutions for improved services (Hoult et al., 2009).

Among these technologies, wireless smart cameras are particularly prominent. Cameras equipped with AI and automation capabilities offer real-time monitoring and anomaly detection in sectors ranging from public transportation to building surveillance and wildlife management. For instance, Jang et al. (2019) employed cameras and line scanning with deep learning to detect anomalies in rail systems, showcasing the efficacy of camera based IoT systems for monitoring, protection, and timely insights. Similar advancements have been made in addressing challenges associated with monitoring surfaces exposed to weather conditions and constant motion, such as roads, to enhance image quality in suboptimal conditions (Chun & Ryu, 2019; Ai & Kwon, 2020).

Despite these advancements, current smart infrastructure initiatives often prioritize human needs, overlooking the potential to benefit wildlife and contribute to a more sustainable urban ecosystem (Azevedo Guedes et al., 2018; Mayrand & Clergeau, 2018). This anthropocentric bias limits the potential of smart cities to address broader ecological concerns and foster coexistence with non-human inhabitants. Our research aims to bridge this gap by leveraging smart infrastructure technologies for wildlife monitoring and conservation, contributing to a more holistic approach to urban infrastructure management.

Urban wildlife and the imperative of monitoring

While much of the current discourse on smart infrastructure centers on human-centric benefits like optimized energy consumption, improved transportation, and enhanced urban living (Neuman, 2020; Popescu, 2022), there is a growing recognition of the need to consider a "more-than-human" perspective. This perspective emphasizes the role of cities as habitats for diverse plant and wildlife species, acknowledging their contribution to human well-being and ecosystem health (Bektas et al., 2018; Maynard and Clergeau, 2018).

Research is increasingly revealing the potential for smart infrastructure to support biodiversity and contribute to sustainable ecosystems (Maynard and Clergeau, 2018; Neuman, 2020; Lashof and Neuberger, 2023). Green spaces, urban forests, and rooftop gardens, for instance, can serve as vital habitats for wildlife while also improving air quality and providing recreational areas for humans

(Neuman, 2020). However, city managers are sometimes focused on how to remove unwanted animals from city infrastructure, rather than promoting coexistence (Hadidian, 2015). Recent research has indicated that there are costs to decreasing biodiversity that are directly linked to human mortality (Frank, 2024). By designing cities and their technologies with all inhabitants in mind, we can foster biodiversity conservation alongside smart infrastructure deployment (Bektas et al., 2018).

Moreover, integrating green elements into infrastructure, such as green walls and roofs, can provide additional living spaces for wildlife and promote sustainable human-computer interaction (Maynard and Clergeau, 2018). This approach aligns with the broader goal of designing technology for sustainability, enabling smart cities to serve all inhabitants efficiently while reducing power consumption, traffic congestion, and pollution (Neuman, 2020; Hansson et al., 2021).

Beyond environmental sustainability, smart cities also have the potential to support both human and non-human interests. Popescu (2022) explores how the management of waste, water, and infrastructure systems can be optimized to reduce waste, improve recycling, and mitigate environmental and infrastructure degradation. These efforts, while often focused on human communities, can unintentionally benefit wildlife by preserving ecosystems and reducing the impact of extreme weather events. Lashof and Neuberger (2023) further highlight the need for adaptive infrastructure that prioritizes both human and environmental health by reducing emissions.

Effective wildlife monitoring is essential for understanding urban ecosystems and making informed decisions about infrastructure development and maintenance. However, traditional monitoring methods, such as visual surveys, often face limitations in terms of accuracy, efficiency, and the ability to collect long-term data (Civjan et al., 2017; Adam & Hayes, 2000; Hendricks et al., 2005). This is particularly true for species like bats, which often roost in less accessible areas like bridges and tunnels.

The challenges associated with wildlife monitoring underscore the need for innovative approaches that leverage technology to provide reliable, longitudinal data. Such data can empower decision-makers to anticipate conflicts, minimize environmental impact, and foster a more sustainable coexistence between humans and wildlife in our cities.

More-than-human perspective in HCI: designing for coexistence

The design and implementation of technology carry significant implications, particularly concerning their impact on the environment and non-human inhabitants. While smart buildings offer energy efficiency and enhanced functionality for human users (Atzori, Iera, & Morabito, 2010), the field of human-computer interaction (HCI) has traditionally been anthropocentric, often overlooking the needs of other species. This gap is particularly evident in research on small mammals like bats, which are underrepresented in HCI and animal-computer interaction (ACI) studies (Koche et al., 2021).

However, the aforementioned 'more-than-human' perspective is emerging within HCI, advocating for a shift away from human-centered design and towards acknowledging the interconnectedness of humans, technology, and the natural world (Vallgårda & Redström, 2017). This approach emphasizes recognizing and valuing the agency and the influence of non-human entities in the design and use of technology. It encourages us to consider how our technological creations impact the environment and other species, fostering a sense of shared responsibility and coexistence.

Furthermore, sustainable human-computer interaction (sHCI) emphasizes the importance of designing technology that is environmentally sustainable, socially responsible, and economically viable in the long term (Biggs et al., 2021; Dourish, 2010; Hansson et al., 2021). This includes

considering the entire lifecycle of technology, from resource extraction to disposal, and ensuring its minimal impact on the environment.

The collection and analysis of real-time data on wildlife activity patterns, enabled by smart infrastructure technologies, can play a crucial role in promoting sustainable coexistence. By understanding how non-human inhabitants utilize urban spaces, we can make more informed decisions about urban planning and infrastructure development. For example, data on bat roosting locations can inform maintenance schedules for bridges, while minimizing disruptions to their habitat. This data-driven approach aligns with the 'more-than-human' perspective by acknowledging the agency of non-human entities and incorporating their needs into the decision-making process.

Ultimately, a shift towards a more-than-human approach in HCI and urban planning can help us create cities that are not only smart and efficient but also sustainable and inclusive for all inhabitants, both human and non-human.

To demonstrate the potential of a more-than-human approach in urban planning and wildlife conservation, we present a case study focused on monitoring bat populations in bridges using smart infrastructure technologies. In the following section, we outline the methodology employed in this study, detailing the data collection, analysis, and visualization techniques used to gain insights into bat activity patterns and inform decision-making.

Methods

Case study approach

This research employed a case study methodology, involving a collaboration between university researchers from the fields of biology and information technology, and the Ohio department of transportation. The primary goal was to develop and evaluate a system for monitoring bat activity in bridges, contributing to informed decision-making regarding infrastructure maintenance and wildlife conservation. The case study involved the deployment of IoT-enabled cameras on bridges to collect data on bat presence and activity patterns. This data was then integrated into a custom-designed, cloud-based dashboard, BatMap, to facilitate visualization and analysis.

Data collection: deploying trail cameras on bridges

To assess the feasibility of using trail cameras for bat detection, we tested several trail cameras for performance in this use case. We selected and strategically deployed Reolink Keen Ranger PT cameras under 27 bridges in Ohio, USA. The camera model was chosen for its rapid trigger speed, no-glow night vision, high video quality, and internet connectivity, features crucial for capturing clear videos of bats, particularly at night.

Deployment sites were selected based on cellular service availability for remote control and data retrieval, while avoiding bridges over high-traffic areas or large waterways. Cameras were mounted in potential bat roosting areas, such as expansion joints (Bektas et al., 2018), using double-sided tape (Figure 1). The Reolink mobile app facilitated precise camera aiming (Figure 2). All fieldwork was conducted with necessary permissions and in coordination with authorities.



Figure 1. A trail camera deployed underneath a bridge

Initial trials revealed improved bat detection and video quality with supplemental infrared (IR) lighting. Therefore, external battery-powered IR illuminators with timer-controlled power supply were deployed at each camera location. Cameras remained active for 1–23 days (average = 10.2 ± 5.5 SD), limited by factors like battery life, theft, or study completion. Captured videos were transmitted to the Reolink cloud service and subsequently downloaded to Microsoft OneDrive for storage and analysis. Two experienced bat ecologists reviewed all videos to confirm bat presence and identify species.

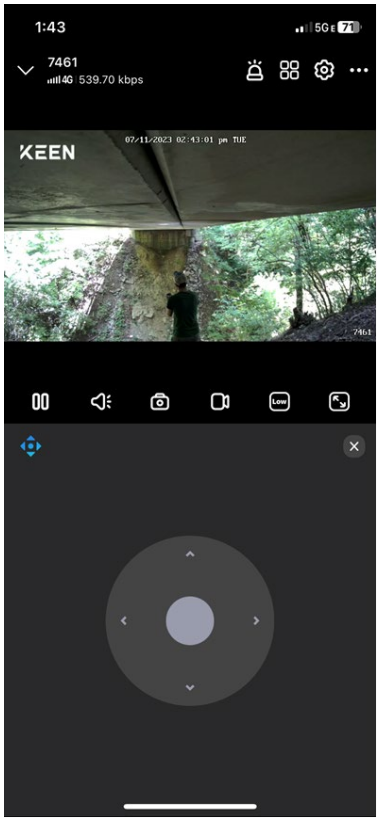


Figure 2. Screen capture from Reolink mobile app, showing researcher aiming a camera during deployment

Data analysis and visualization: BatMap dashboard

To effectively manage and visualize the collected data, we developed BatMap, a map-based dashboard tailored to the department of transportation's needs. It utilized HTML, CSS, and JavaScript for the front end and a MySQL database on a university web server for the back end, handling various requests.

BatMap's ability to display data across time allows for insights into long-term bat activity patterns, addressing limitations of traditional surveys and accounting for bat migration. It aims to facilitate informed decision-making regarding infrastructure projects and minimize disturbances to bat populations.

This case study approach, combining IoT-enabled cameras with a custom-designed dashboard, allows us to investigate the effectiveness of this system for bat monitoring (RQ1) and evaluate how the dashboard interface supports decision-making related to infrastructure and conservation (RQ2).

Results

Effectiveness of IoT-enabled cameras for bat detection and monitoring (RQ1)

Our deployment of IoT-enabled cameras across 27 bridges resulted in bat detection at 4 sites (15%), offering initial evidence for the potential of this technology in transforming bridges into smart infrastructure for wildlife monitoring. Over a 60-day deployment period at these occupied sites, bats were recorded on 8 nights (13%). The captured videos clearly showed bats entering and exiting the bridges, typically within a few seconds, demonstrating the system's ability to detect and record bat activity (Figure 3).

These findings suggest that while the overall detection rate across all deployed sites was relatively low, the system proved effective in identifying bat presence at bridges where they were actively roosting or utilizing the structures. The high-quality video capture, particularly at night, facilitated clear identification of bats despite their rapid movements.

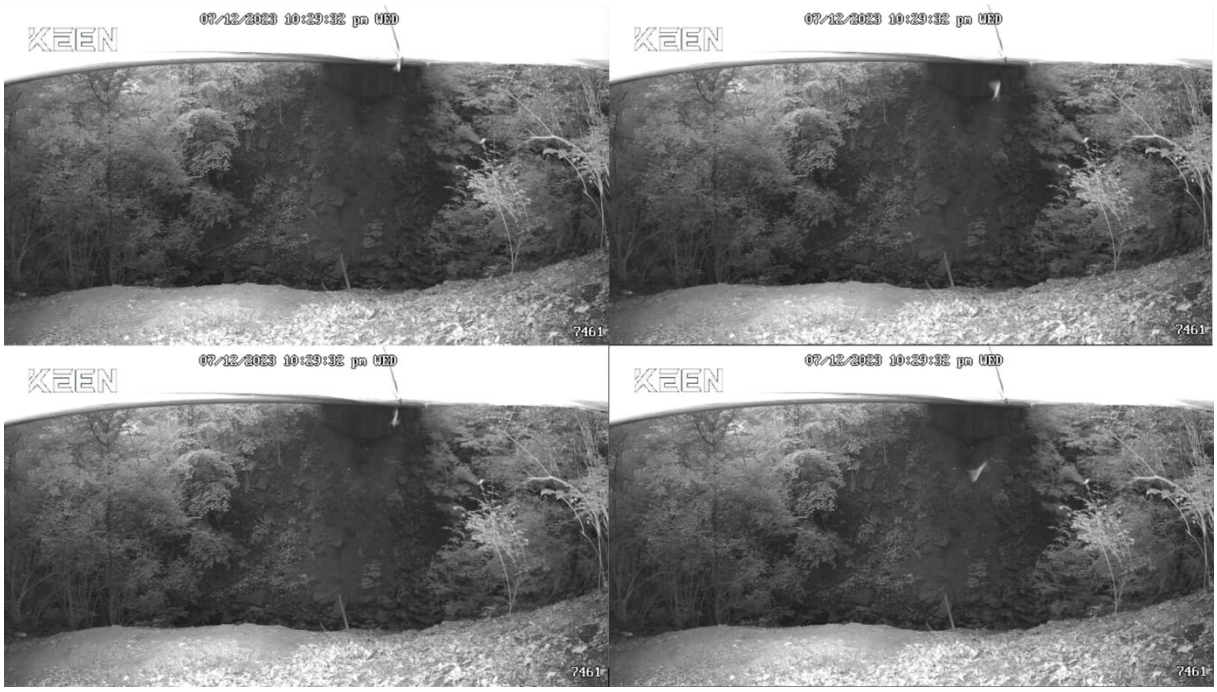


Figure 3. Bat capture evidence from our camera deployment

BatMap dashboard: supporting decision-making (RQ2)

BatMap's interface was designed to facilitate informed decision-making regarding infrastructure maintenance and wildlife conservation by providing readily accessible and relevant information to users. All features were discussed with collaborators at the state Department of Transportation during monthly meetings prior to implementation. The following describes features that were included in the design. This is followed by a description of how the system was designed to meet the decision-making needs of the Department of Transportation, and an example of how the dashboard has been used.

Split-screen layout: the map and search/local inventory side-by-side configuration allows users to simultaneously view spatial data on bridge locations, natural areas, and waterways, fostering a holistic understanding of the ecological context surrounding infrastructure (Figure 4).

Bridge List: this feature offers high-level information on bridges displayed on the map, including the Structure File Number (SFN), district, county, and any existing survey data (Figure 4). This enables project planners to quickly identify bridges requiring bat surveys or prioritize maintenance based on proximity to bat habitats or other critical natural resources.

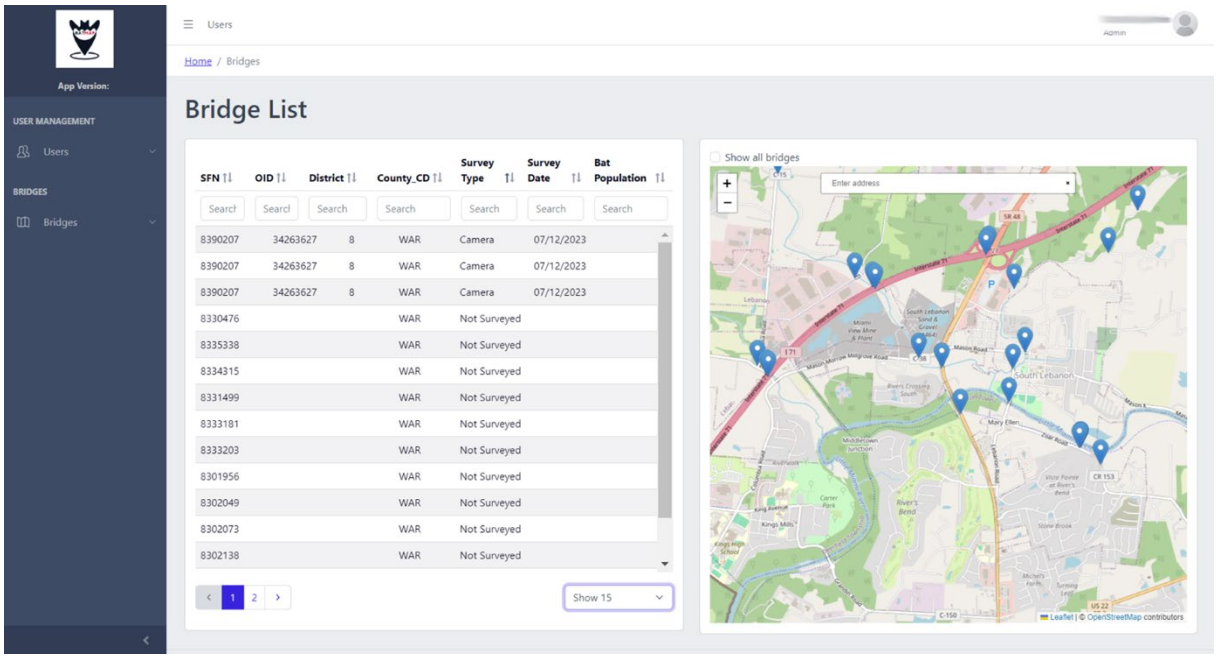


Figure 4. Screen capture of a user's split-screen view upon logging in to BatMap

Bridge Profile: clicking on a bridge in the bridge list or on the map opens a detailed profile displaying structural attributes and environmental factors known to influence bat use. This feature caters to both experienced bat ecologists and those less familiar with bats, thus aiding in informed decision-making (Figure 5). The inclusion of uploaded photos and videos provides visual evidence of bat presence, further aiding in assessment.

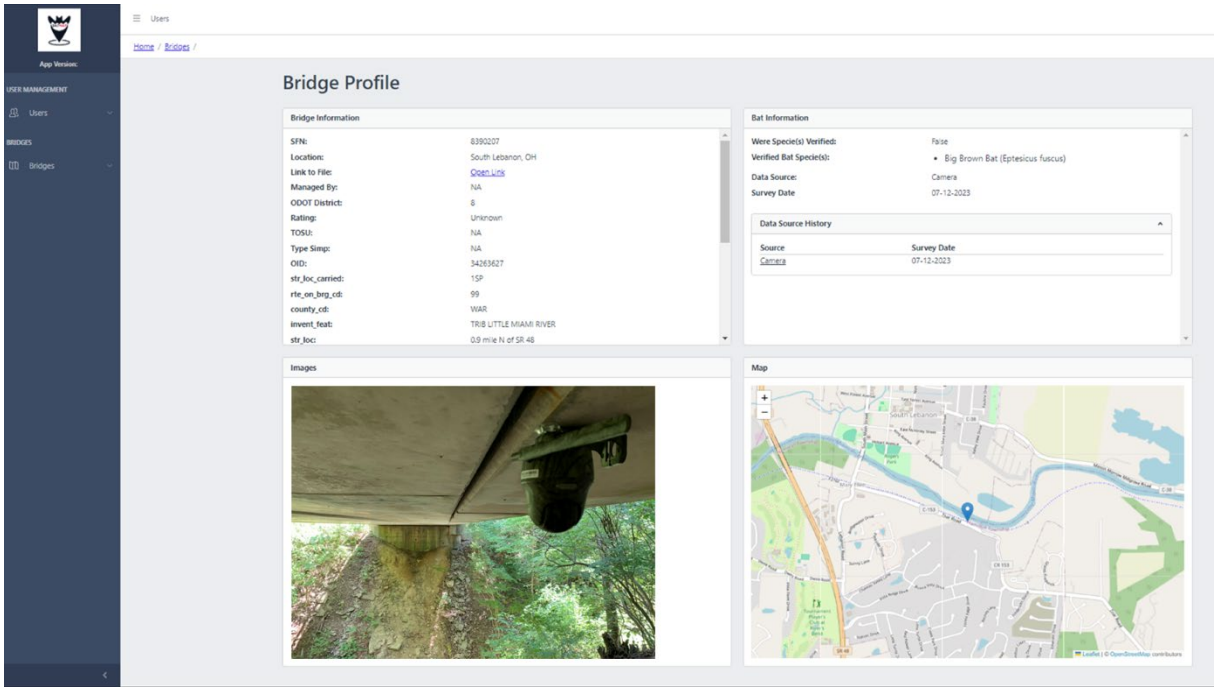


Figure 5. Screen capture of BatMap's Bridge Profile page

Map Integration: the persistent map within the Bridge Profile reinforces the spatial context, allowing users to visualize the bridge's location in relation to surrounding natural features.

By integrating these features, BatMap supports decision-making processes related to infrastructure maintenance and wildlife conservation in several ways:

Prioritization of Surveys: the system enables project planners to efficiently identify bridges in need of bat surveys, prioritizing those with no prior data, proximity to known bat habitats, or areas of ecological significance.

Informed Maintenance Planning: the Bridge Profile provides insights into the likelihood of bat presence based on structural and environmental attributes, enabling planners to schedule maintenance activities during periods of low bat activity or implement mitigation measures to minimize disturbance.

Data-Driven Conservation Efforts: the longitudinal data collected by the system can contribute to broader conservation efforts by identifying critical bat habitats and informing the design of wildlife-friendly infrastructure.

Knowledge Sharing and Awareness: BatMap's user-friendly interface and accessible information foster knowledge sharing and raise awareness about the importance of considering wildlife in infrastructure planning and management. The system was built with a level of security to ensure that the locations of sensitive species are not publicly available but are accessible to relevant managers who require such information for conservation and decision-making purposes.

In summary, BatMap's design is intended to empower decision-makers with the information and tools needed to balance infrastructure development with wildlife conservation while ensuring the safety of sensitive information, fostering a more sustainable coexistence between humans and the natural world in urban environments. In the same year that we made BatMap available to the Department of Transportation, the dashboard was used effectively to respond to a report of a colony of state-endangered bats roosting underneath a bridge that was scheduled for replacement

in several years. Although the planning and contracting for the bridge replacement had already been completed, BatMap was still used to acquire situation awareness. Specifically, Department of Transportation planners were able to query the bridge from the Bridge List and see that no bat surveys had ever been conducted at the site. They also used the integrated map to quickly determine if surveys had been conducted at any nearby bridges to determine if additional conflicts between planning and conservation might exist. Finally, they were able to devise a plan for mitigating impacts on the bat colony by considering the ecological context of the bridge, including where supplemental habitat could be provided for the bats during construction.

Discussion

The results of this study offer several design implications for the development of smart infrastructure systems that support both human needs and wildlife conservation.

Effectiveness of IoT-enabled cameras

The successful detection of bats at 15% of deployed sites demonstrates the potential of IoT-enabled cameras as a tool for wildlife monitoring, aligns with research that emphasizes the importance of careful deployment strategies to maximize detection rates in wildlife monitoring using camera traps (Rovero & Zimmermann, 2016). However, the low overall detection rate suggests the need for further refinement and optimization of camera placement and settings. Future research could explore the use of machine learning algorithms to improve bat detection accuracy and reduce false positives (e.g., recordings of insects), echoing studies that have successfully applied AI-powered image analysis to identify and classify various wildlife species (Norouzzadeh et al., 2018).

BatMap dashboard for decision-making

The positive feedback from the Department of Transportation collaborators regarding BatMap's interface and features underscores its potential to support decision-making processes, highlighting the importance of user-centered design in developing effective decision-support tools (Kujala, 2003). The split-screen layout, bridge list, and bridge profile were thought to be particularly useful for prioritizing surveys, planning maintenance activities, and fostering awareness of bat presence. These design elements, particularly the combination of spatial and attribute-based information presentation, align with research on geospatial visualization, which emphasizes the value of integrating multiple data layers to facilitate understanding and decision-making (MacEachren & Kraak, 2001). By providing users with both spatial and attribute-based information in an easily accessible and understandable format, BatMap empowers decision-makers to make informed choices that balance infrastructure development with the needs of wildlife.

More-than-human design

The integration of wildlife monitoring capabilities into smart infrastructure reflects the growing interest in "more-than-human" design in HCI, which challenges anthropocentric perspectives and advocates for considering the needs of non-human entities (Vallgårda & Redström, 2017). This study demonstrates how such an approach can lead to developing systems that not only enhance human experiences but also contribute to environmental sustainability and wildlife conservation, aligning with the concept of sustainable HCI, which emphasizes the need for technology to support both social and ecological well-being (Blevis, 2007).

Data-driven conservation

The collection of longitudinal data on bat activity patterns has the potential to inform broader conservation efforts and contribute to the design of wildlife-friendly infrastructure. This aligns with research emphasizing the importance of long-term monitoring for understanding wildlife populations and their responses to environmental change (Magurran et al., 2010). Future research

could explore the use of this data to model bat habitat preferences and predict potential conflicts with infrastructure development, echoing studies that employ species distribution modeling to inform conservation planning and mitigate human-wildlife conflicts (Guisan et al., 2013).

Limitations and future directions

While this study provides promising results, it is important to acknowledge certain limitations. The relatively small sample size of bridges in one mid-western US state and the focus on a single taxonomic group (bats) limits the generalizability of the findings. This resonates with calls for more inclusive and diverse perspectives in HCI research, including a broader consideration of non-human stakeholders (Bardzell, 2010). Further research is needed to explore the applicability of this approach to other wildlife species and infrastructure types.

Additionally, the study relied on commercially available trail cameras, which may have limitations in terms of sensitivity and image quality in certain conditions. One such limitation is the ability of the built-in IR lights, to illuminate small, fast-flying bats from distances >20 m. The development of specialized sensors or the integration of AI-powered image analysis could further enhance the system's capabilities. This aligns with research on the use of emerging technologies, such as drones and bioacoustics, in ecological research and conservation (Linchant et al., 2015; Aide et al., 2013).

Future research should also investigate the long-term impact of this system on decision-making processes and conservation outcomes. It would be valuable to assess whether the data provided by BatMap leads to tangible changes in infrastructure maintenance practices and contributes to the protection of bat populations. Overall, this study demonstrates the potential of integrating wildlife monitoring capabilities into smart infrastructure to support a more-than-human approach to urban planning and development. By leveraging technology and data-driven insights, we can create cities that are not only efficient and livable for humans but also sustainable and inclusive for all inhabitants.

Conclusion

This study demonstrates the potential of integrating IoT-enabled cameras and a custom-designed dashboard, BatMap, for effective bat monitoring and informed decision-making in the context of urban infrastructure maintenance. Our study successfully demonstrates the potential of IoT-enabled cameras for bat monitoring (RQ1). While bat activity was detected at a moderate 15% of surveyed bridges, the system's ability to capture clear video evidence, even in low-light conditions, underscores its promise for transforming bridges into valuable components of smart infrastructure for wildlife conservation. These findings suggest that with further refinements, such as optimizing camera placement and incorporating AI-powered detection algorithms, the effectiveness of this approach can be significantly enhanced.

Moreover, the positive reception of BatMap by Department of Transportation collaborators highlights the value of a user-friendly, map-based dashboard in supporting decision-making (RQ2). By integrating spatial data visualization, detailed bridge profiles, and multimedia evidence, BatMap empowers users to prioritize surveys, plan maintenance activities strategically, and foster awareness of bat presence. This demonstrates the system's potential to facilitate data-driven decision-making that balances infrastructure development with the needs of wildlife, contributing to a more sustainable coexistence in urban environments.

This research contributes to the fields of information and data science, particularly within the context of the iConference community. We showcase a novel application of IoT technology and data visualization in the realm of wildlife monitoring and conservation, underscoring the valuable potential of interdisciplinary collaboration between information technology and biology.

Furthermore, our work champions a more-than-human approach by integrating wildlife monitoring capabilities into smart infrastructure, thereby adding to the growing body of research that challenges anthropocentric design paradigms. This approach resonates with the iSchools' Climate Action Coalition's vision, promoting interdisciplinary efforts towards climate resilience and sustainable coexistence.

BatMap's user-centric design empowers decision-makers with accessible and pertinent information, enabling them to make informed choices that minimize the environmental impact of infrastructure development. This exemplifies the power of data science in bridging the gap between research and real-world application, fostering evidence-based decision-making that benefits both human and non-human communities.

In conclusion, our study highlights the feasibility and positive impact of employing IoT-enabled cameras and a tailored dashboard for bat monitoring and conservation within urban settings. These findings emphasize the necessity of adopting a more-than-human perspective in the design and implementation of smart infrastructure, ultimately envisioning cities that are not only technologically advanced but also ecologically balanced and sustainable.

Acknowledgements

This work was made possible by a grant from the Ohio Department of Transportation's STAR program.

About the authors

Jacob Koch is an Adjunct Professor and Ph.D. candidate in the School of Information Technology at the University of Cincinnati. He has a B.A. in History from the University of Cincinnati, with a focus in 19th century American history. He has a master's degree in Computer Information Technology from Northern Kentucky University. His research interests are primarily in cloud computing and virtualization technologies, as well as IoT, and cybersecurity.

Jess Kropczynski is an Associate Professor of Information Technology at the University of Cincinnati. She focuses on designing and evaluating technology that helps communities work together. Her research explores how people use technology for collective action in areas like city planning and emergency response.

Janette Perez-Jimenez is a Ph.D. candidate in the Department of Biological Sciences at the University of Cincinnati. Her research focuses on the conservation biology of bats in Ohio.

Joseph S Johnson is an Assistant Professor of Information Technology in the School of Information Technology at the University of Cincinnati. He holds a Ph.D. in Animal Sciences and a master's degree in Forestry from the University of Kentucky. His research on devising and applying evidence-based solutions for animal conservation.

References

Adam, M. D., & Hayes, J. P. (2000). Use of bridges as night roosts by bats in the Oregon Coast Range. *Journal of Mammalogy*, *81*(2), 402-407.

Ai S., Kwon J. Extreme low-light image enhancement for surveillance cameras using attention U-net. Sensors. 2020;20:495. doi: 10.3390/s20020495

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1, e103.

Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15), 2787-2805. https://doi.org/10.1016/j.comnet.2010.05.010

Azevedo Guedes, A. L., Carvalho Alvarenga, J., Dos Santos Sgarbi Goulart, M., Rodriguez y Rodriguez, M. V., & Pereira Soares, C. A. (2018). Smart Cities: The Main Drivers for Increasing the Intelligence of Cities. *Sustainability*, 10(9), 3121. https://doi.org/10.3390/su10093121

Bardzell, S. (2010). Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1301-1310).

Bektas, B. A., Hans, Z., Phares, B., Nketah, E., Carey, J., Solberg, M. K., & McPeek, K. (2018). Most Likely Bridges as Roosting Habitat for Bats: Study for Iowa. Transportation Research Record, 2672(24), 1-10. https://doi.org/10.1177/0361198118758649

Biggs, H., Joshi, T., Murphy, R., Bardzell, J., & Bardzell, S. (2021). Alternatives to agrilogistics: Designing for ecological thinking. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-31. https://doi.org/10.1145/3491102.3491239

Blevis, E. (2007). Sustainable interaction design: Invention & disposal, renewal & reuse. In Proceedings of the *SIGCHI conference on Human factors in computing systems* (pp. 503-512).

Chun C., Ryu S.-K. Road surface damage detection using fully convolutional neural networks and semi-supervised learning. Sensors. 2019;19:5501.

Civjan, S. A., Berthaume, A., Bennett, A., & Dumont, E. (2017). Bat Roosting in Bridges: Pros and Cons of Assessment Methods from a New England Regional Study. *Transportation Research Record*, 2628(1), 120-128. https://doi.org/10.3141/2628-13

Daniel Lashof and Jillian Neuberger, Climate-smart infrastructure in the United States—what does it look like and how do we get it built? , 2023 Environ. Res.: Infrastruct. Sustain. 3 012003

Dourish, P. (2010, August). HCI and environmental sustainability: The politics of design and the design of politics. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems* (pp. 1-10). Forman, R. T. T. (2014). Urban Ecology: Science of Cities. Cambridge University Press.

Frank, E. G. (2024). The economic impacts of ecosystem disruptions: Costs from substituting biological pest control. *Science*, 385, eadg0344. https://doi.org/10.1126/science.adg0344

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., & Martin, T.G. (2013). Predicting species distributions for conservation decisions. Ecology letters, 16(12), 1424-1435.

Hadidian, J. (2015). Wildlife in U.S. Cities: Managing unwanted animals. *Animals (Basel)*, 5(4), 1092–1113. https://doi.org/10.3390/ani5040401

Hansson, L. Å. E. J., Cerratto Pargman, T., & Pargman, D. S. (2021, May). A decade of sustainable HCI: Connecting SHCI to the sustainable development goals. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–19).

Hendricks, P., Lenard, S., Currier, C., & Johnson, J. (2005). *Bat use of highway bridges in south-central Montana* (Report to Montana Department of Transportation). Montana Natural Heritage Program, Helena. 31 pp.

Hoult, Neil, Bennett, Peter J., Stoianov, Ivan, Fidler, Paul, Maksimović, Čedo, Middleton, Campbell, Graham, Nigel, Soga, Kenichi, Wireless sensor networks: creating 'smart infrastructure', 2009, Proceedings of the Institution of Civil Engineers - Civil Engineering, 136-143 doi: 10.1680/cien.2009.162.3.136

iSchools. (n.d.). Climate Action Coalition. Retrieved September 13, 2024, from https://www.ischools.org/climate-action-coalition

Jang J., Shin M., Lim S., Park J., Kim J., Paik J. Intelligent image-based railway inspection system using deep learning-based object detection and weber contrast-based image comparison. Sensors. 2019; 19:4738.

Kujala, S. (2003). User-centered design. In *The human-computer interaction handbook* (pp. 100-115). CRC Press.

Linchant, J., Lisein, J., Semeki, J., Lejeune, P., & Vermeulen, C. (2015). Are unmanned aerial vehicles (UAVs) the future of wildlife monitoring? A review of accomplishments and challenges. *Mammal Review*, 45(4), 239-252.

MacEachren, A. M., & Kraak, M. J. (2001). Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1), 3-12.

Magurran, A.E., Baillie, S.R., Buckland, S.T., Dick, J.M., Elston, D.A., Scott, E.M., Smith, R.I., Somerfield, P.J., & Watt, A.D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. Trends in Ecology & Evolution, 25(10), 574-582.

Mayrand, F., & Clergeau, P. (2018). Green roofs and green walls for biodiversity conservation: A contribution to urban connectivity? *Sustainability*, 10(4), 985. https://doi.org/10.3390/su10040985

Neuman, M. Infrastructure Is Key to Make Cities Sustainable. *Sustainability* **2020**, *12*, 8308. https://doi.org/10.3390/su12208308

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the *National Academy of Sciences*, 115(25), E5716-E5725.

Ogie, Robert Ighodaro, Perez, Pascal, Dignum, Virginia, Smart infrastructure: an emerging frontier for multidisciplinary research, 2017, Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction, 8-16 Doi:10.1680/jsmic.16.00002

Popescu, S. (2022). Towards Sustainable Urban Futures: Exploring Environmental Initiatives in Smart Cities. *Applied Research in Artificial Intelligence and Cloud Computing*, 5(1), 84–104. Retrieved from https://www.researchberg.com/index.php/araic/article/view/130

Rejeb, A., Rejeb, K., Simske, S., Treiblmaier, H., & Zailani, S. (2022). The big picture on the internet of things and the smart city: A review of what we know and what we need to know. Internet of Things, 19, 100565. https://doi.org/10.1016/j.iot.2022.100565

Rovero, F., & Zimmermann, F. (2016). *Camera trapping for wildlife research*. Pelagic Publishing.

Vallgårda, A., & Redström, J. (2017). From human-centered to more-than-human design. *interactions*, 24(6), 42-45.

Wilson, C., Hargreaves, T., & Hauxwell-Baldwin, R. (2015). Smart homes and their users: A systematic analysis and key challenges. *Personal and Ubiquitous Computing*, 19(5), 463–476. https://doi.org/10.1007/s00779-014-0813-0

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Improving scholarship accessibility with reinforcement learning

Haining Wang, Jason Clark, Hannah McKelvey, Leila Sterman, Gao Zheng, Zuoyu Tian, and Xiaozhong Liu

DOI: https://doi.org/10.47989/ir30iConf47530

Abstract

Introduction. A vast amount of scholarly work is published daily, yet much of it remains inaccessible to the general public due to dense jargon and complex language. We introduce a reinforcement learning approach that fine-tunes a language model to rewrite scholarly abstracts into more comprehensible versions.

Method. Our approach utilises a carefully balanced combination of word- and sentence-level accessibility rewards to guide the language model in substituting technical terms with more accessible alternatives, a task which models supervised fine-tuned or guided by conventional readability measures struggle to accomplish.

Analysis. We evaluate our model's performance through readability metrics, factual accuracy assessments and language quality measurements, comparing results against supervised fine-tuning baselines.

Results. Our best model adjusts the readability level of scholarly abstracts by approximately six US grade levels—in other words, from a postgraduate to a high school level. This translates to roughly a 90% relative improvement over the supervised fine-tuning baseline, while maintaining factual accuracy and high-quality language.

Conclusion. We envision our work as a step toward bridging the gap between scholarly research and the general public, particularly younger readers, and those without a college degree.

Introduction

Accessible language in science communication

At first glance, the daily publication of tens of thousands of scientific papers—many freely accessible through open science and open access initiatives—suggests few barriers to knowledge dissemination. However, two key facts challenge this perception and show that significant barriers remain. A recent survey by the US Department of Education found that more than half of US adults aged 16 to 74 (54%, or 130 million people) read at or below a sixth-grade level (Rothwell, 2020). Meanwhile, an analysis of the readability of biomedical research abstracts published from 1881 to 2015 found that scientific writing has become increasingly difficult to read over time (Plavén-Sigray et al., 2017). Even when intended to be accessible, scientific abstracts typically require a postgraduate level of reading comprehension due to jargon use and sentence structure (Wang and Clark, 2024). This discrepancy leaves a significant portion of the population—including young readers and adults without advanced degrees unable to fully engage with scientific works, even if these are made freely available online. The '*infodemic*' surrounding COVID-19 highlighted this issue: the urgent need for understandable information about the virus clashed with the complex presentation of scientific findings—leading many to turn to more digestible but less reliable narratives on social media (Wang et al., 2019; Islam et al., 2020; Calleja et al., 2021).

While the legal and medical fields have long been encouraged to use accessible language as a clear conduit for public engagement (Mazur, 2000; Petelin, 2010), momentum for the adoption of accessible language within scientific communities has been building, roughly since the start of the open science movement (Schriver, 2017). For instance, the National Institutes of Health (NIH) advocates for 'clear and simple' principles when communicating with audiences with limited health literacy, and the *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) requires authors to submit a significance statement accessible to non-experts (Berenbaum, 2021; Pool et al., 2021). However, there are inherent conflicts between the specialised nature of communication among disciplinary peer scholars and the public-oriented dissemination of scientific findings. Even assuming that communicating scholarly works in plain language is possible, it will inevitably increase the communication cost among domain experts and create confusion at the more advanced levels, compared to the use of jargon and technical terms. In an era of increasingly specialized scientific research, this conundrum is not easily addressed by scientists or disseminators. Hence, given the current landscape, the widespread adoption of accessible language in scholarly works is unlikely in the near future.

In response, we propose addressing the need for communicating scientific findings to a broader audience by *rewriting scholarly abstracts with simpler words and grammar using language models.* Since readability is key to comprehending scholarship (Flesch, 1946; DuBay, 2004; Kerwer et al., 2021), we envision the resulting accessible narratives as paving the way for the 'last mile' of science, broadening access to scientific understanding and engagement, especially for younger readers and those without a university degree.

Challenges to effective simplification

Fine-tuning a language model using pairs of abstracts and their accessible versions is the *de facto* method for automating the rewriting of scholarly abstracts into more accessible versions (Xu et al., 2015; Goldsack et al., 2022; Joseph et al., 2023). Accordingly, we introduced the Scientific Abstract-Significance Statement (SASS) corpus (Wang & Clark, 2024), a dataset composed of paired abstracts and significance statements from diverse disciplines, with the latter targeting 'an undergraduate-educated scientist outside their field of specialty' (Berenbaum, 2021; Pool et al., 2021). Although the simplified abstracts generated from language models fine-tuned on the SASS corpus are approximately three grade levels more readable than the original abstracts, as measured by US grade-based readability scores (Wang & Clark, 2024, Sec. 6), the documents are still not sufficiently accessible; even the best models produce college-level texts. Additionally,

because the vocabulary used in significance statements is often just as complex as that found in the abstracts themselves (Wang & Clark, 2024, Sec. 3), the readability improvements are primarily due to shorter sentences, and technical terms remain inadequately addressed.

Alternatively, the optimisation of a language model can be guided by a chosen objective in an actor-critic manner (Ramamurthy et al., 2023). It is intuitive to choose an established document readability measure, such as the Automated Readability Index (ARI; see Section 4.2.2), to assess the overall readability of the outputs generated by the language model. However, we found that the optimisation of language models guided by ARI is highly unstable, often resulting in the production of seemingly more accessible versions that still contain many technical terms. Inspired by Riddell and Igarashi (2021), we decomposed the measurement of document readability into two distinct measures: one at the sentence level and one at the word level. We then prioritized word-level accessibility in the optimisation to encourage the model to use more accessible words instead of simply shortening sentences.

Contribution

Our work aims to serve as a bridge between scholarly works and the general public, particularly benefiting younger readers and those without a college degree.

1. We address the common challenges in science communication by rewriting scholarly abstracts at a high school reading level using a language model.

2. We identify the challenges language models face in properly addressing jargon and propose Reinforcement Learning from Accessibility Measures (RLAM) as a means to improve the models' use of accessible terms in their rewrites. RLAM-trained language models can significantly reduce the reading level of a scholarly abstract from a postgraduate level to a high school level, achieving a 3 grade-level reduction or about a 90% performance boost compared to models fine-tuned using the same corpus.

3. We observe systematic differences between reinforcement learning models guided by different rewards and conclude that disproportionate weights for sentence-level rewards contribute to unstable training and lower simplification quality.

Our code, model generations and training logs are available at https://github.com/ Wang-Haining/RLAM under a permissive licence.

Scientific abstract-significance statement (SASS) corpus

We used the scientific abstract significance statement (SASS) corpus in our experiments. This corpus is composed of 3,430 abstract-significance statement pairs derived from PNAS and divided into training (3,030 samples), validation (200 samples), and test sets (200 samples) (Wang & Clark, 2024). It covers a wide range of disciplines, ensuring diverse representation across various fields, as shown in Figure 1. Corpus statistics are shown in Table 1; refer to Section 4.2 for a detailed description of the measures.

Section	ARI ↓	F-K ↓	VOA	SL ↓	WA	WL ↓
Abstract	18.9 (2.8)	19.2 (2.4)	-0.43 (0.25)	25.4 (4.9)	12.0 (0.4)	5.3 (0.4)
Significance	18.1* (3.1)	18.6* (2.7)	-0.31* (0.26)	23.9* (5.3)	11.9* (0.4)	5.4* (0.4)

Table 1. Corpus statistics for the scientific abstract-significance statement (SASS) corpus. Metrics include ARI (Automated Readability Index), F-K (Flesch-Kincaid readability test), VOA (log ratio of proportion of words found in the VOA1500 vocabulary), SL (average sentence length and number of sentences), WA (word accessibility; log frequency per 1 billion tokens in English Wikipedia), and WL (average word length). Measures whose names are followed by a down arrow symbol (↓) indicate that lower values correspond to a more readable document. Numeric values in parentheses are the corresponding standard deviations. Paired t-tests were conducted for each metric comparing the abstracts and significance statements, with p-values adjusted using the Bonferroni correction for multiple comparisons. The observed differences in each of the measurements are statistically significant after adjusting for the grouped p-values at a significance level of 0.05.

We observed that significance statements are semantically coherent with their corresponding abstracts. The corpus statistics indicate that significance statements are more readable than abstracts, as shown by lower mean values in the Automated Readability Index (ARI) and Flesch-Kincaid readability test (F-K). This suggests that the SASS corpus can be useful in simplifying scholarly abstracts across diverse disciplines.

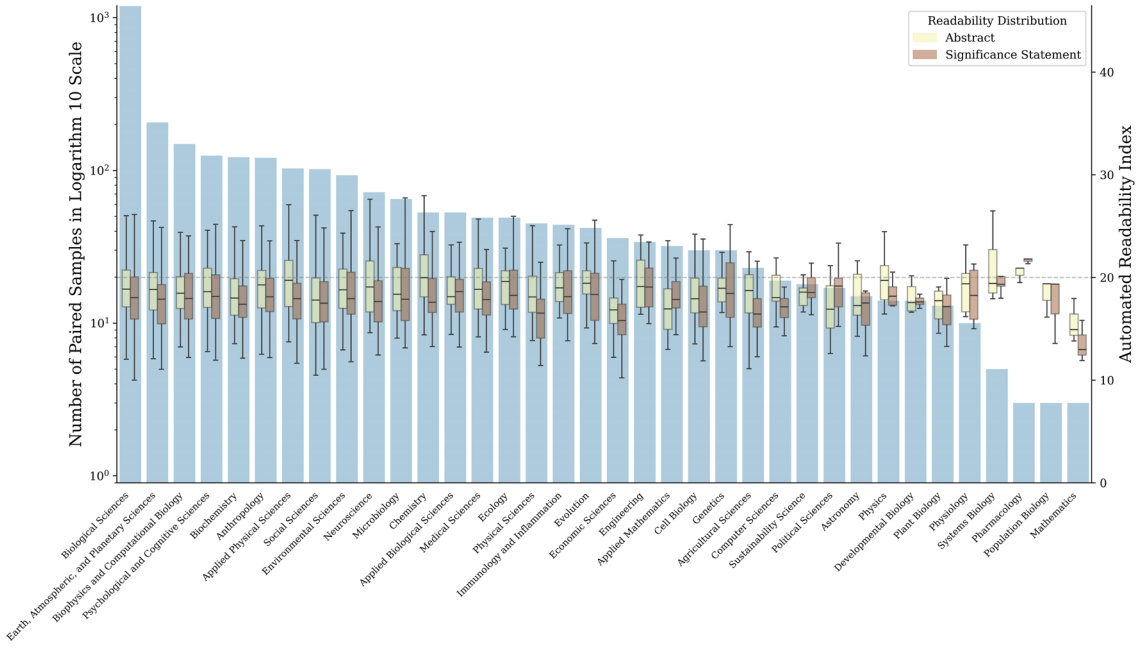


Figure 1. Discipline and readability distributions of abstracts and significance statements found in the training set of the Scientific Abstract-Significance Statement corpus. The count of paired samples in different disciplines is shown in blue bars on a log10 scale (disciplines with fewer than three samples are not shown). Readability is measured using the Automated Readability Index (ARI), which estimates the number of years of schooling required to understand a text. On average, abstracts have a readability slightly below 20 ARI, indicating a post-graduate level. Significance statements are generally more readable than their corresponding abstracts.

We also observed that word accessibility (i.e., log frequency per 1 billion tokens found in English Wikipedia) and average word length suggest that significance statements can be less accessible at

the word level than are their corresponding abstracts. Although the log ratio of words found in the VOA1500 vocabulary is slightly lower than in the corresponding abstracts, these 1,500 words are very basic and include a high proportion of function words. Considering that significance statements use approximately 1.5 fewer words on average, the increased use of VOA words may be a consequence of the higher use of function words to maintain grammaticality.

Reinforcement learning from accessibility measures

Language modelling via proximal policy optimisation

At the core of our approach is language modelling with Proximal Policy Optimisation (PPO) (Schulman et al., 2017) guided by two accessibility measures. A causal language model trained on large corpora can generate the next token based on the current sequence, which is useful in the context of reinforcement learning for developing a policy model that determines the most appropriate next token to maximize the expected return in terms of document readability.

The process begins with an input sequence s0 = (a_0, a_1,..., ai), where each a_i is from

a set of tokens W, and s_0 represents an abstract formatted in a simple template. The language model π_θ then generates a_0, a_1,..., aT_{-1} ⊏ π_θ ($\cdot \mid s_t$), creating its accessible version until the maximum number of tokens T is reached, either due to the context length or an end-of-sentence token:

$$\pi_\theta(a_0, a_1, \ldots, a_{T-1}) = \prod_{t=0}^{T-1} \pi_\theta\,(a_t|s_t) \tag{1}$$

Our objective is to learn a policy model that, given an abstract, models the joint probability of tokens leading to a high reward in terms of accessibility while maintaining semantic coherence. Formally, this is expressed as:

$$J(\pi\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}(r(s_t, a_t) - \beta_{\mathrm{KL}}\mathrm{KL}(\pi\theta(a_t|\, s_t) \parallel \pi_{\theta_{\mathrm{SFT}}}(a_t|\, s_t)))\right] \tag{2}$$

Here, $J(\pi_\theta)$ represents the expected return when following policy π_θ. The reward $r(s_t, at)$ is estimated for each time step t in the trajectory τ = (s_0,a_0,s_1,a_1,..., sT_{-1},a_{T-1}), where s_t is the sequence of tokens at time t, formed as the concatenation of s_0 and the tokens a_0,a_1,...,a_{t-1}. This formula iteratively computes the rewards given the current sequence s_t and the token a_t chosen by the policy. The β_{KL}-weighted KL divergence term KL($\pi_\theta(a_t \mid s_t)$ // $\pi_{\theta\mathrm{SFT}}(a_t \mid s_t)$) is applied at every step of sequence generation to ensure the policy does not deviate significantly from the supervised fine-tuned model. This is crucial because, without such a constraint, the policy model might quickly learn to output whatever the reward model favours to maximize its return, which can lead to undesirable behaviours. For instance, the model might repeatedly output a frequent word ('is is is ...'), achieving a high reward based solely on accessibility measures but lacking meaningful content. Following Stiennon et al. (2020), β_{KL} is dynamically adjusted by targeting a specific KL divergence between π_θ and $\pi_{\theta\mathrm{SFT}}$ using a capped proportional controller in logarithmic space. The benefit of using a dynamic KL control, as opposed to a fixed one, is that it allows the model to adapt more flexibly to different stages of training, accommodating varying levels of KL divergence between the policy and SFT model.

In practice, we fine-tune the policy model in an actor-critic manner: while the policy model (the actor) generates sequences of tokens based on the current sequence s_t, the critic is an additional linear layer that takes the output of the language model's last layer and produces a scalar for time step t, estimating the expected cumulative reward of producing the token a_t, noted as $V_\rho(a_t)$. The problem is reduced to optimising at every step the expected cumulative reward of taking action a_t

in state s_t and following the policy π_θ thereafter ($V_\rho(s_{t+1}) = V_\rho(s_t,a_t)$), compared to the expected cumulative reward of being in state s ($V_\rho(s_t)$), termed as the advantage $d_t = V_\rho(s_{t+1}) - V_\rho(s_t)$.

We used the final reward of the entire generation (see Section 3.3) and back-propagate it through the sequence using Temporal Difference (TD) and Generalised Advantage Estimation (GAE) to estimate the advantage of each token:

$$d_t = \sum_{t'=t}^{T} (\gamma\lambda)^{t'-t} \left(r_T + \gamma V_\rho(s_{t'+1}) - V_\rho(s_{t'})\right) \quad (3)$$

where γ is the discount factor for future rewards, λ controls the bias-variance trade-off, and r_T is the final reward, which, in our case, is a linear combination of two accessibility measures. V_ρ is trained by minimizing the square error loss; see Equation 5.

We used the Proximal Policy Optimisation (PPO) (Schulman et al., 2017) clipped surrogate objective with importance sampling to more efficiently use offline samples to update the online policy. Importance sampling corrects for the discrepancy between the behaviour policy that generated the samples and the current policy by weighting the samples using the ratio of their probabilities under both policies. The PPO algorithm introduces a clipping mechanism to balance exploration and exploitation while preventing large, potentially harmful updates to the policy, see Equation 4. The whole RLAM algorithm is illustrated in Algorithm 1.

Algorithm 1 Training with Reinforcement Learning from Uncombined Accessibility Measures. The policy model is updated using the PPO clipped surrogate objective (Eq. 4), and the value model is updated by minimising a square-error objective (Eq. 5).

1: **Input:** initial policy model $\pi_{\theta_{SFT}}$, randomly initiated value head $V_{\rho_{init}}$, final reward function r_T for the last token, weighted by β_{KL} for KL divergence term; task prompts X; hyperparameters $\gamma,\lambda,\square$

2: $\pi_\theta \leftarrow \theta_{SFT}$, $V_\rho \leftarrow \rho_{init}$

3: **for** step = 1,..., M **do**

4: Sample a batch $\{s_0\}^n$ from X

5: Sample output sequences $\{a_0, a_1,..., a_{T-1}\}^n \square \pi_\theta(\cdot \mid s_0)$ for each prompt s_0 in the batch ▷ Eq. 1

6: Compute final reward r_T for each sampled output sequence $\{a_0, a_1,...,a_{T-1}\}^n$. ▷Sec. 3.3

7: Distribute the final reward r_T to each token in the sequence through GAE ▷ Eq. 3

8: Compute advantages $\{d_t|s_t\}_{t=0}^{T-1}$, value targets $\{V_{targ}(s_t)\}_{t=0}^{T-1}$ for each sequence with V_ρ and compute KL divergence penalty $KL_t = KL(\pi_\theta(a_t \mid s_t) \,//\, \pi_{\theta_{SFT}}(a_t \mid s_t))$

9: **for** PPO iteration = 1,..., μ **do**

10: Update the policy model by maximizing the PPO clipped surrogate objective with KL penalty:

$$\theta \leftarrow \arg\max_{\theta} \frac{1}{n}\sum_{n=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\min\left(\frac{\pi_{\theta_{\text{online}}}(a_t|s_t)}{\pi_{\theta_{\text{offline}}}(a_t|s_t)}(A_t - \beta_{\text{KL}}\text{KL}_t),\right.$$
$$\left.\text{clip}\left(\frac{\pi_{\theta_{\text{online}}}(a_t|s_t)}{\pi_{\theta_{\text{offline}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon\right)(A_t - \beta_{\text{KL}}\text{KL}_t)\right) \quad (4)$$

11: **end for**

12: Update the value model by minimizing a square-error objective:

$$\rho \leftarrow \arg\min_{\rho} \frac{1}{n}\sum_{n=1}^{n}\frac{1}{T}\sum_{t=1}^{T}(V_\rho(s_t) - V_{\text{targ}}(s_t))^2 \quad (5)$$

13: **end for**

14: **Output:** π_θ

Adaptive Kullback–Leibler controller

Following Ziegler et al. (2019, Sec. 2.2), we dynamically adjust β_{KL} to target a specific KL divergence value, $\text{KL}_{\text{target}}$, using a log-space proportional controller.

The update rule is:

$$\beta_{\text{KL}_{t+1}} = \beta_{\text{KL}_t}\left(\left(1 + K_\beta.\,\text{clip}\left(\frac{\text{KL}(\pi_\theta,\ \pi_{\theta_{\text{SFT}}}) - \text{KL}_{\text{target}}}{\text{KL}_{\text{target}}}, -0.2, 0.2\right)\right)\right)$$

where K_β is the proportional gain, set to 0.01.

Reward function

The reward function evaluates the overall quality of the output (s_T). After the initial failures of testing a traditional readability measure (i.e., ARI) as the criterion, we decided to use a balance of two accessibility measures: average sentence length in words and word accessibility, adopted from Riddell and Igarashi (2021), to guide the optimisation.

Word accessibility reward A word's accessibility is approximated by how frequently it appears in a large reference corpus. We chose the English Wikipedia corpus due to its domain similarity and applied a Moses tokenizer, yielding a vocabulary of 14.6 million types from a total of 3.6 billion tokens. If a token is among the most common 100,000 types, we report its frequency per billion tokens as its accessibility measure. Otherwise, we estimate its frequency using ridge regression with an $l2$-norm coefficient equal to 1.0. This model allows us to make serviceable estimates of the frequency of arbitrary tokens, including tokens that do not appear in the reference corpus. This model takes as input the token's length in Unicode code points, its byte unigrams, byte bigrams, and byte trigrams. The model estimates the token's log frequency per 1 billion tokens. We used the natural logarithm of frequencies per billion tokens as the measure of word accessibility. For example, the accessibility score for 'big' is 11.8, while 'colossal' scores 7.3. Despite being comparable in meaning, the model's production of the latter will receive fewer rewards. Coefficient β_{wa} is to control the scale of the credit given for word accessibility.

We have faithfully followed the experiment of Riddell and Igarashi (2021) with three differences. First, our reference corpus is the English Wikipedia, whereas the original study used the Common Crawl News corpus. Second, we did not discard duplicated sentences as Riddell and Igarashi (2021) did, because we found that sentence duplication is not common in Wikipedia. Third, the original study reported word *inaccessibility* scores by negating the logarithm of frequency per billion. We report *accessibility*, without negation, because it is more naturally suited to serve as a reward. Refer to Riddell and Igarashi (2021, pp. 1186–1187) for the training of the ridge regression.

Sentence length reward Sentence length is also determined by a Moses tokeniser, which preserves hyphenation and splits contractions. For the Moses rule-based tokeniser, we use the *sacremoses* Python package. We negate the value of sentence length for intuitive calculation of the rewards for optimisation.

Experiment setup

Training

We initialised the policy models π_θ by adopting the Gemma-2B checkpoint reported in Wang and Clark (2024) ($\pi_{\theta_{SFT}}$). The original Gemma-2B was trained on three trillion tokens, consisting of publicly available data as well as proprietary datasets comprising 'primarily English data from web documents' (Mesnard et al., 2024). The specific checkpoint we adopted was fine-tuned using the SASS corpus in a straightforward manner. It was chosen for its strong performance for simplification quality, faithfulness, and relatively compact size.

The two accessibility rewards were weighted as follows: the word accessibility reward was set to β_{WA} = 4.0, and we report two models with different β_{SL} values of 0.05 and 0.2. Because word accessibility is in logarithmic space, we subtracted 10 from the average word accessibility of the output and reset any values lower than 10 to 0 to keep them within a reasonable range. For adaptive control of the per-token semantic reward, we started with an initial β_{KL} = 0.2 and targeted a KL divergence of 8.0 nats during the training course, capping it in the range between 0.15 and 0.25. We used a micro batch size of 4, with each sequence used to run the PPO algorithm for 4 epochs using importance sampling, with gradient accumulation steps set to 4. We used a clip range of 0.2 for the policy gradient and value function estimation to ensure stability. The value function coefficient was set to 0.1. The optimisation used standard AdamW optimiser parameters (β_1 = 0.9, β_2 = 0.999, $\varepsilon = 1 \times 10^{-8}$) (Kingma and Ba, 2015; Loshchilov and Hutter, 2017). The learning rate was fixed at 1×10^{-6}. The training was conducted on 2 H100 (80GB) GPUs using mixed precision training in bfloat16. The sampling temperature was set to 0.7 in the rollout phase. Following Huang et al. (2024), we assigned unfinished rollouts (indicated by the missing end of sequence token) with a fixed low score to encourage the model to generate complete simplified narratives. We also report a model trained as dictated by ARI (i.e., RLARI; described in Section 1.2), with all other parameters kept the same as those guided by the accessibility measures. We tested other hyperparameters for guiding the optimisation of RLARI in pilot studies but obtained similar results.

We performed multiple runs for each reinforcement learning process and selected the checkpoint to report based on a balance of semantic retention and ARI score obtained on the validation set. We observed that the readability of the generated text on the validation set often began to decline rapidly after plateauing for a while, typically accompanied by a surge in the standard deviation of average word accessibility among sentences. This signals that the language model was sacrificing language quality and semantic relevance for extra improvements in readability. Therefore, we reported the checkpoint immediately before such instability occurred.

Evaluation

We evaluated the simplified texts generated by the language models trained with the reinforcement learning framework using 200 abstracts from the SASS corpus test set for

simplification. Though advanced decoding methods might further refine the quality of the outputs, we used multinomial sampling with the temperature set to zero to intentionally produce the most deterministic outputs. This approach helped us better understand the modelling of accessible language and made it easier for us to identify potential quirks. We assessed the quality of the generated simplified abstracts both quantitatively and qualitatively. Quantitatively, we measured the generated texts based on their semantic retention and accessibility using established relevance measures as well as readability and accessibility metrics.

Semantic retention

BERTScore calculates the cosine similarity between each token in the candidate sentence and each token in the reference sentence using contextual embeddings from a pre-trained language model; its results align well with human judgement on semantic similarity evaluation (Zhang et al., 2019). It is not directly influenced by lexical overlap, making it more suitable for evaluating simplification systems than are metrics that rely on matching words, such as BLEU (Papineni et al., 2002). For our evaluation, we used embeddings from the 18th layer of a BERT-large-uncased model and reported the F1 score. This choice is based on prior findings indicating that the 18th layer yields a strong Pearson correlation (0.72) on the WMT16 To-English benchmark (Zhang et al., 2019).

Simplification and accessibility

Accessibility can be measured with respect to the overall simplification quality (SARI); readability (ARI and Flesch-Kincaid); and other straightforward document complexity measures, including average sentence length, word accessibility (i.e., log frequency per 1 billion tokens found in English Wikipedia), the log ratio of its proportion of VOA Special English words (1,517 types in total), and average word length.

SARI (System output Against References and against the Input sentence) is specifically designed to evaluate text simplification (Xu et al., 2016). It aims to measure how well a simplified text retains the original meaning while improving readability. SARI provides a balanced measure of how well a text simplification system performs by focusing on the necessary operations of adding, deleting, and retaining words.

ARI and Flesch-Kincaid readability tests assign a numerical score to text that reflects the US grade level required for comprehension. Lower scores (1–13) indicate content suitable for kindergarten through twelfth grade, with each score corresponding to a subsequent grade level. Scores in the range of 14–18 suggest college-level readability, ranging from first- to senior-year content appropriateness. Higher scores (19 and above) are associated with advanced college education. Both measures use average sentence length. Flesch-Kincaid uses syllables per word, while ARI uses characters per word for its linear combination with sentence length.

We harvested VOA Special English vocabulary comprising 1,517 unique words (VOA1500). We included VOA Special English Word Book Sections A-Z, Science Programs, and Organs of the Body hosted on Wikipedia (Wikimedia Foundation, 2024). We calculate the ratio of words that appear in the VOA1500 to those that do not, then report the natural logarithm of this ratio for each generated sample. Values above 0 indicate that the text contains more Special English words than non-Special English words, and a higher value indicates a greater presence of ‘easy’ words.

Language quality, faithfulness, and completeness

We manually examined 5% of all generated samples, corresponding to a randomly chosen subset of the test set from the SASS corpus. Each generation is annotated with respect to language quality, faithfulness, and completeness using a rubric of Good, Acceptable, and Poor. We focused on fluency and grammaticality and hand-picked both good and problematic examples when evaluating language quality. For the evaluation of faithfulness, we conduct close readings to assess the extent to which a simplified abstract remains factually faithful to the original narrative. If

uncertainty arises, we consult the corresponding manuscript, as our abstract simplification system must avoid producing misinformation. Completeness is also a key consideration, as it is essential to include the main findings and implications of the research for the general public, since this is the primary goal of scientific dissemination.

Findings and discussion

Quantitative assessment

Table 2 summarises the performance of Gemma-2B, tuned in different ways, when evaluated on the test set of the SASS corpus. The first scenario is the supervised fine-tuned baseline (SFT), which performs next-token prediction on the SASS corpus training set.

The second and third scenarios are reinforcement learning through PPO guided by ARI (RLARI) or accessibility measures (RLAM). We assessed the generation quality by considering both semantic retention and simplification, specifically using BERT score (BS), SARI, ARI, Flesch-Kincaid readability test (F-K), the log ratio of words in the VOA1500 vocabulary (VOA), sentence length (SL), word accessibility (WA), and word length (WL). A one-tailed paired t-test was conducted for each metric to compare observations between the reinforcement learning and supervised fine-tuning baselines, assuming improvement in document readability. Bonferroni correction was applied to each set of tests to maintain a family-wise significance level of 0.05.

Model	β_{SL}	β_{WA}	ARI ↓	F-K ↓	SARI	VOA	SL ↓	WA	WL ↓	BS
SFT	-	-	15.5 (3.0)	16.5 (2.6)	39.1 (5.0)	-0.26 (0.30)	20.6 (4.1)	11.9 (0.5)	5.2 (0.4)	0.64 (0.06)
RLARI	-	-	12.6* (2.9)	14.3* (2.5)	40.1* (4.8)	-0.17* (0.31)	16.4* (3.7)	12.0 (0.5)	5.0* (0.4)	0.64 (0.05)
RLAM	0.05	4.0	13.5* (2.8)	14.8* (2.4)	39.8 (5.1)	0.08* (0.29)	21.0 (4.2)	12.7* (0.5)	4.8* (0.4)	0.62 (0.06)
RLAM	0.2	4.0	12.5* (2.9)	14.0* (2.5)	39.8 (5.0)	-0.01* (0.32)	17.7* (3.4)	12.4* (0.5)	4.9* (0.4)	0.63 (0.05)

Table 2. Comparison of Gemma-2B's performance across three approaches: the supervised fine-tuned baseline (SFT), reinforcement learning guided by ARI (using an intermediate checkpoint before significant policy gradient instability was observed, RLARI), and reinforcement learning guided by two accessibility measures (RLAM). SFT was fine-tuned using the Scientific Abstract-Significance Statement (SASS) corpus. The columns labelled β_{WA} and β_{SL} pertain specifically to RLAM, where the rewards for average word accessibility and sentence length are balanced. The inference on the test split from SASS uses multinomial sampling. Metrics ARI, F-K, SARI, VOA, SL, WA, WL, and BS stand for Automated

Readability Index, Flesch-Kincaid readability test, log ratio of VOA1500 vocabulary, sentence length, word accessibility, word length, and BERTScore (F1), respectively. Measures followed by a down arrow symbol (↓) indicate that lower values are better. Numeric values in parentheses are the corresponding standard deviations. A paired two-tailed t-test was performed on observations of each measure between each model and the original abstracts. At a model-wise p-value of 0.05, measures that differ significantly from the SFT baseline are marked with an asterisk.

We observe that reinforcement learning models trained with different rewards exhibit a notable reduction in reading level, bringing abstracts down to high school levels. The model directly guided by ARI (RLARI) achieves an ARI of 12.6, while the most performant model guided by accessibility measures (RLAM, β_{SL} = 4.0 and β_{WA} = 0.2) reaches 12.5, both aligning with the readability level expected for individuals who have completed K-12 education (approximately ARI 13). However,

RLARI and RLAM models achieve these readability improvements in different ways. For RLAM models, better readability is achieved mostly through improved token-level accessibility. RLAM models show an increase in word accessibility from 0.5 to 0.8 compared to the supervised baseline. This increase in the natural logarithm suggests that words generated by RLAM models are, on average, 1.6 to 2.2 times more frequent in the English Wikipedia corpus than those generated by the SFT model. In comparison, RLARI's 0.1-unit increase in word accessibility, although observed, does not result in a statistically significant change in word frequency compared to the SFT model. Similarly, the log ratio of the VOA1500 vocabulary in the RLAM models shows a significant improvement, with log ratios ranging from −0.02 to

0.08. This implies that for every 100 non-VOA1500 (or more complex) words generated, RLAM models can produce approximately 99 to 106 VOA1500 basic words. In contrast, the SFT and RLARI models exhibit VOA log ratios of −0.26 and −0.17, respectively, indicating that for every 100 non-VOA1500 words generated, these models produce only around 77 (84) VOA1500 words. The average word length in characters for RLAM models ranges from 4.8 to 4.9, slightly shorter than RLARI's and outperforming SFT. Overall, the above evidence suggests that RLAM models achieve better readability by using more common, simpler, and shorter words.

On the other hand, the RLARI model achieves better readability by producing much shorter sentences, with only a marginal boost in word-level accessibility. The RLARI model has the shortest average sentence length of 16.4 words, significantly outperforming the SFT model. In comparison, a significantly shorter average sentence length is only observed when RLAM's sentence length reward coefficient (β_{SL}) exceeds 0.08. We also observed that, in pilot studies, if β_{SL} is set to a higher value, such as 0.5, the model's optimisation will collapse after only a few hundred steps, similar to what we consistently observed in the optimisation of RLARI models. The improvement in the RLARI model's word-level accessibility is inconsistent: while we observed significant gains in the VOA log ratio and word length, the word accessibility measure did not show statistically significant improvement after a Bonferroni correction. That said, although RLARI uses more VOA basic words and shorter words, their frequency in the English Wikipedia corpus is not high enough to result in a significant increase in word accessibility compared to the SFT baseline.

For semantic coherence with the human-written significance statements, the BERT scores from the reinforcement learning models are not significantly worse than those of the SFT model. This finding suggests that the generations from the reinforcement learning models are likely to remain semantically faithful and that the reinforcement learning process does not significantly degrade language quality, at least up to the point before unstable optimisation signals were observed. SARI scores from the reinforcement learning models are significantly, yet only marginally, better than those of the SFT baseline. Since SARI is a measurement of a combination of deletions, additions, and retention operations, a straightforward explanation is not available. However, the higher SARI scores confirm the greater simplification quality of the reinforcement learning models, which is likely due to a combination of simpler words and shorter sentences.

Qualitative analysis

We annotated 5% of the generated simplified abstracts from reinforcement learning models guided by ARI (RLARI) and accessibility measures (RLAM, with $\beta_{WA} = 4.0$ and varying β_{SL} values) with respect to language quality, faithfulness, and completeness, as shown in Figure 2. The test abstracts included three from biological sciences and one each from chemistry; mathematics; evolutionary biology; environmental sciences; ecology; economic sciences; and earth, atmospheric, and planetary sciences.

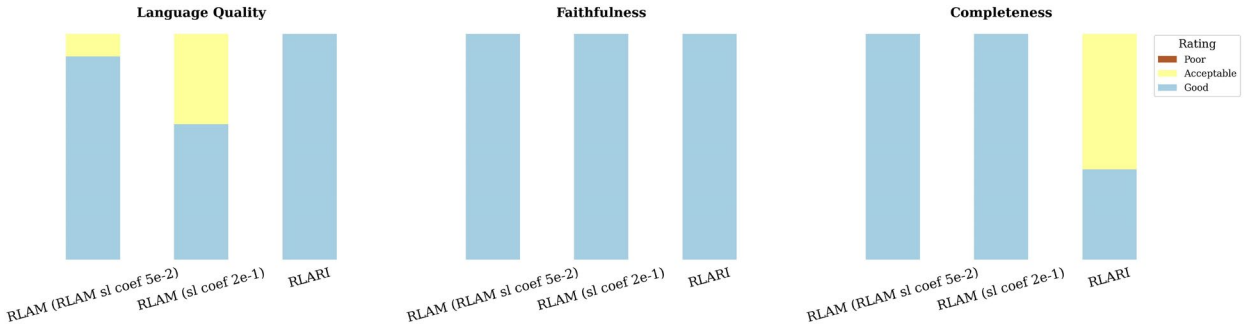


Figure 2. We annotated 5% of the generated outputs from reinforcement learning models guided by ARI (RLARI) and accessibility measures (RLAM, with $\beta_{WA} = 4.0$ and two different β_{SL} values).

Regarding overall quality, we found that reinforcement learning-trained models generally produced high-quality language. Compared to the SFT generations, RLAM outputs are often shorter and more semantically complete, due to the imposed token budget for newly generated content (241 tokens, the length of the longest significance statement in the training set). The main shortcoming is the presence of small trailing phrases that often deflate readability scores, such as '(PsycINFO Database Record),' '(show more),' and 'All rights reserved.' The first artefact is also found in the SFT model generations and is hypothesised to be carried over from the corpora that Gemma-2B was previously exposed to, as this pattern does not appear in the SASS corpus. An informal review of the remaining generations suggests that this phenomenon is amplified, appearing even in samples without such trailers in the SFT generations. The latter two artefacts were newly observed and are caused by the reinforcement learning processes. Where any of these issues appear, we annotate them as 'Acceptable,' even though the generations are otherwise fluent and grammatical. We also found that the proportion of these artefacts steadily increases as training continues, and they are usually found in only a subset of samples across different experiments.

In assessments of faithfulness, we did not find any models hallucinating in the generations we examined. However, in informal examinations, we did find that reinforcement learning checkpoints may hallucinate by generating simple but overly hedging or short expressions when over-optimised. Generations from RLARI-trained models often remain unfinished, but they retain the main gist of the abstract. Although we do not observe trailer phrases in the RLARI-generated texts like those found in the RLAM models, this issue frequently arises in other RLARI runs. Subsequent checkpoints often exhibit similar problems and tend to deteriorate rapidly once the model starts cutting corners. This typically occurs shortly before the training process completely fails. However, the reported checkpoint happens to miss this characteristic.

Conclusion

To improve the accessibility of scientific literature to the general public, we implemented reinforcement learning techniques to guide language models, extending beyond the traditional cross-entropy objective. Our study demonstrates that carefully balancing accessibility measures at the word and sentence levels can effectively guide Gemma-2B in simplifying scholarly abstracts, outperforming the supervised fine-tuning baseline by a large margin. This approach achieves these improvements without compromising language quality or faithfulness and mitigates the supervised fine-tuning model's tendency to overemphasise research implications. The best model trained using our method successfully adjusts the readability level of scholarly abstracts by approximately six US grade levels in other words, from a postgraduate to a high school level. Compared to the supervised fine-tuning model, the words generated by the model trained via our

approach are proven to be more common (1.6 to 2.2 times more frequent), easier (with more VOA basic words), and shorter (by 0.3 to 0.4 characters). This improvement addresses a key limitation of existing corpora, in which the target distribution (i.e., significance statements) often does not adequately prioritise the accessibility of word choice. We hope this work contributes to bridging the gap between scholarship and a broader audience, advancing the understanding and development of better simplification systems, and ultimately fostering a more informed and engaged society.

Acknowledgements

We gratefully acknowledge the support of the Institute of Museum and Library Services (No. RE-246450-OLS-20) and the National Social Science Fund of China (No. 23&ZD221). We also thank the organisers of the LIS Education and Data Science Integrated Network Group (LEADING), including Jane Greenberg, Erik Mitchell, Kenning Arlitsch, Jonathan Wheeler, and Samantha Grabus. Additionally, we are thankful to Coltran Hophan-Nichols and Alexander Salois from the University Information Technology Research Cyberinfrastructure at Montana State University for providing computational resources on the Tempest High Performance Computing System, Doralyn Rossmann for research support, and Deanna Zarrillo for early involvement in the project.

About the authors

Haining Wang is a doctoral candidate in Information Science at Indiana University Bloomington. His research focuses on natural language processing with applications in the humanities, social sciences, and biomedical sciences. He can be reached at hw56@iu.edu.

Jason A. Clark is a Professor and Head of Research Optimisation, Analytics, and Data Services (ROADS) at Montana State University Library. His research interests include machine learning, digital libraries, and the intersection of artificial intelligence with user experience, including algorithmic literacy and machine learning patterns. He can be contacted at jaclark@montana.edu.

Hannah McKelvey is an Associate Professor at Montana State University Library. Her research interests are focused on practice-based librarianship, focusing on electronic resource management, user discovery behaviour, and collection assessment through usage analytics. She can be contacted at hannah.mckelvey@montana.edu.

Leila Sterman is an Associate Professor at Montana State University Library. Her research interests focus on scholarly communication, including open access publishing, copyright, and institutional repositories. She can be contacted at leila.sterman@montana.edu.

Zheng Gao is currently a senior algorithm engineer at Ant Group. His research interests encompass machine learning, deep learning, and large language models, including LLM post-training and evaluation. He can be contacted at gao27@alumni.iu.edu.

Zuoyu Tian is an Assistant Professor at Macalester College. His research interests lie in computational linguistics and natural language processing, with a particular focus on applying computationally intensive methods to study language variation and change. He can be contacted at ztian@macalester.edu.

Xiaozhong Liu is an Associate Professor in Computer Science and Data Science at Worcester Polytechnic Institute. His research interests include natural language processing, text/graph mining, information retrieval/recommendation, metadata, and computational social science. He can be contacted at xliu14@wpi.edu.

References

Berenbaum, M. R. (2021). On COVID-19, cognitive bias, and open access.

Calleja, N., AbdAllah, A., Abad, N., Ahmed, N., Albarracin, D., Altieri, E., Anoko, J. N., Arcos, R., Azlan, A. A., Bayer, J., Bechmann, A., Bezbaruah, S., Briand, S. C., Brooks, I., Bucci, L. M., Burzo, S., Czerniak, C., De Domenico, M., Dunn, A. G., Ecker, U. K. H., Espinosa, L., Francois, C., Gradon, K., Gruzd, A., Gülgün, B. S., Haydarov, R., Hurley, C., Astuti, S. I., Ishizumi, A., Johnson, N., Johnson Restrepo, D., Kajimoto, M., Koyuncu, A., Kulkarni, S., Lamichhane, J., Lewis, R., Mahajan, A., Mandil, A., McAweeney, E., Messer, M., Moy, W., Ndumbi Ngamala, P., Nguyen, T., Nunn, M., Omer, S. B., Pagliari, C., Patel, P., Phuong, L., Prybylski, D., Rashidian, A., Rempel, E., Rubinelli, S., Sacco, P., Schneider, A., Shu, K., Smith, M., Sufehmi, H., Tangcharoensathien, V., Terry, R., Thacker, N., Trewinnard, T., Turner, S., Tworek, H., Uakkas, S., Vraga, E., Wardle, C., Wasserman, H., Wilhelm, E., Würz, A., Yau, B., Zhou, L., and Purnat, T. D. (2021). A public health research agenda for managing infodemics: Methods and results of the first who infodemiology conference. *JMIR Infodemiology*, 1(1): e30979.

DuBay, W. H. (2004). The principles of readability. Technical report, Impact Information, Costa Mesa, CA.

Flesch, R. (1946). *The Art of Plain Talk*. Harper & Row, New York, first edition.

Goldsack, T., Zhang, Z., Lin, C., and Scarton, C. (2022). Making science simple: Corpora for the lay summarisation of scientific literature. *arXiv preprint arXiv*:2210.09932.

Huang, S., Noukhovitch, M., Hosseini, A., Rasul, K., Wang, W., and Tunstall, L. (2024). The N+ implementation details of RLHF with PPO: A case study on TL; DR summarization. arXiv preprint arXiv:2403.17031.

Islam, A. N., Laato, S., Talukder, S., and Sutinen, E. (2020). Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. Technological forecasting and social change, 159:120201.

Joseph, S., Kazanas, K., Reina, K., Ramanathan, V. J., Xu, W., Wallace, B. C., and Li, J. J. (2023). Multilingual simplification of medical texts. arXiv preprint arXiv:2305.12532.

Kerwer, M., Chasiotis, A., Stricker, J., Günther, A., and Rosman, T. (2021). Straight From the Scientist's Mouth—Plain Language Summaries Promote Laypeople's Comprehension and Knowledge Acquisition When Reading About Individual Research Findings in Psychology. Collabra: Psychology, 7(1):18898.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. In International Conference on Learning Representations.

Mazur, B. (2000). Revisiting plain language. Technical communication, 47(2):205–205.

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Petelin, R. (2010). Considering plain language: Issues and initiatives. Corporate Communications: An International Journal, 15(2):205–216.

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., and Thompson, W. H. (2017). Research: The readability of scientific texts is decreasing over time. eLife, 6: e27725.

Pool, J., Fatehi, F., and Akhlaghpour, S. (2021). Infodemic, misinformation and disinformation in pandemics: Scientific landscape and the road ahead for public health informatics research. In Public Health and Informatics, pages 764–768. IOS Press.

Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., and Choi, Y. (2023). Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In The Eleventh International Conference on Learning Representations.

Riddell, A. and Igarashi, Y. (2021). Varieties of plain language. In Mitkov, R. and Angelova, G., editors, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 1180–1187.

Rothwell, J. (2020). Assessing the economic gains of eradicating illiteracy nationally and regionally in the United States. Technical report, Barbara Bush Foundation for Family Literacy and Gallup. Accessed: 2024-07-12.

Schriver, K. A. (2017). Plain language in the US gains momentum: 1940–2015. IEEE Transactions on Professional Communication, 60(4):343–383.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33:3008–3021.

Wang, H. and Clark, J. (2024). Simplifying scholarly abstracts for accessible digital libraries using language models. In 2024 ACM/IEEE Joint Conference on Digital Libraries, JCDL '24, page 8, Hong Kong, China. ACM.

Wang, Y., McKee, M., Torbica, A., and Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. Social science & medicine, 240:112552.

Wikimedia Foundation (2024). VOA Special English Word Book. [Online; accessed 9January-2024].

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: new data can help. Transactions of the Association for Computational Linguistics, 3:283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics, 4:401–415.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

How do data authors perform in data-intensive research activities? Evidence from author contribution statement in data papers

Yang Heng , Yu Yonglin, and Liu Fenghong

DOI: https://doi.org/10.47989/ir30iConf47524

Abstract

Introduction. Despite the increasing prevalence of data-intensive scientific research, the division of labor in these activities and the performance of data authors remain underexplored. By employing the Contributor Roles Taxonomy (CRediT), this study examines the division of scientific labor in data papers from *Data in Brief*.

Method and analysis. Utilizing methods of mathematical statistics and data visualization, we analysed the connections between the 14 CRediT roles within data papers. We also explored the relationship between the distribution of labor and the size and discipline of the authorial team, as well as the associations between key authors and their respective CRediT roles.

Results. The results show that 1) data papers rarely make full use of the 14 CRediT roles to describe author contributions. 2) Team size and discipline have a significant impact on the labor division of data-intensive scientific research activities. 3) The need for data collection and analysis is the main reason for the expansion of team size, which is particularly evident in the natural sciences. 4) Corresponding authors and first authors continue to take on core roles. 5) Meanwhile, undertaking data analysis and processing-related tasks, such as '*Software*', helps authors advance in the author order of data papers.

Conclusion. This study provides insights into the division of labor in data-intensive scientific research and shows that CRediT has limitations in fully capturing the research workflow of data papers. We propose developing a taxonomy specific to data papers, such as DP - CRediT.

Introduction

Scientific data, as a crucial research output, has now widely gained recognition for its academic and economic value (Greenberg, Wu, et al., 2023; Pasquetto et al., 2017; Wilson et al., 2014). Under the impetus of open science, the forms of data publication are becoming increasingly diverse and the content is continually enriched (Landi et al., 2020; Wittenburg, 2021). As the infrastructure for data openness and sharing is progressively established and refined (Benhamed et al., 2023; Greenberg, McClellan, et al., 2023), the issue of how to incentivize researchers' willingness to open their data has become a new challenge for scientific data sharing in an open environment **(**Faniel & Jacobsen, 2010**;** Tenopir et al., 2011**;** Treadway et al., 2016**)**.

In recent years, it has been recognized that appropriate and meaningful incentives are essential to capitalize on the promise of data sharing (Lo & DeMets, 2016) and that crediting data generators is key in this effort (Kalager et al., 2016). As stated by the International Council for Science (ICSU), '*Scientists should be recognized and given credit for the scientific contribution of the data sets that they produce as well as for the analysis of those data*' (ICSU). Consequently, Barbara E. et al. proposed the designation of '*data authors*' as an incentive for data sharing, with explicit identification in publications (Bierer et al., 2017). Notably, more than a decade ago, Jillian C.'s exploratory study found that for many participants, the term '*author*' in the context of data was not fitting, raising the question: according to scientific researchers, is data something that can be authored (Wallis & Borgman, 2011)? We believe that this shift in perception is related to the development of data publication and the rise of data papers over the past decade. In the dissemination of data papers, in order to be cited as a data author, a person must have made substantial contributions to the original acquisition, quality control, and curation of the data, be accountable for all aspects of the accuracy and integrity of the data provided, and ensure that the available data set follows FAIR Guiding Principles (Bierer et al., 2017). However, accurately defining data responsibilities to clarify the identity of data authors is a current challenge.

In data-intensive scientific research activities, establishing the identity of data authors requires that their contributions to the data are quantifiable and evaluable. With the proliferation of author contribution statements, various contributor role ontologies and taxonomies (CROTs) have emerged (Hosseini, Colomb, et al., 2023), offering standardized lists of roles or terms to designate individuals' contributions to research. Among these, CRediT (NISO), as a standardized method for describing author contributions, has been vigorously promoted by numerous journal publishers and widely adopted since its introduction in 2014. It is also gradually being applied to data papers to describe the contributions of authors in the production and publication process of scientific data.

In this research, we employ CRediT to examine the division of scientific labor in a sample of data papers from *Data in Brief*, exploring how research contributions are allocated in data-intensive scientific activities. More specifically, we first explore the intercorrelations among the 14 CRediT roles to assess the utilization of CRediT in data papers. We also consider the relationship between the division of scientific labor and the size of the author team, as well as the discipline of the research. Finally, we investigate the correlation between key authors in data papers, such as corresponding authors and first authors, and the CRediT roles they undertake.

Related work

Invisible labor in data-intensive science

As the fourth paradigm (Nielsen, 2009; Tolle et al., 2011), data-intensive science has been widely discussed in the academic community. Ramachandran et al. regard data-intensive science as a scientific discovery process that is driven by knowledge extracted from large volumes of data rather than the traditional hypothesis-driven discovery process, and they introduce the concept of '*data prospecting*' to address the challenges of data-intensive science (Ramachandran et al.,

2013). Data prospecting requires more interdisciplinary collaboration, Cheruvelil and Soranno argue that data-intensive science will be most successful when used in combination with open science and team science (Cheruvelil & Soranno, 2018).

The division of labor in data-intensive science diverges significantly from traditional research models (Pietsch, 2015), primarily due to the sheer scale and complexity of data management and analysis. Unlike research activities that may be more individualistic or limited in scope, data-intensive science often requires collaborative efforts across multiple disciplines and relies heavily on technological infrastructures (Lenhardt et al., 2016; Schultes et al., 2022). This collaborative aspect introduces a new dimension to the scientific process, where the labor is not just intellectual, but also deeply intertwined with the technical and logistical support systems that enable data collection, processing, and interpretation at massive scales.

Scroggins et al. apply the concept of invisible labor to data-intensive science (Scroggins & Pasquetto, 2020). Drawing on a fifteen-year corpus of research into multiple domains of data-intensive science, they used a series of ethnographic vignettes to offer a snapshot of the varieties and valences of labor in data-intensive science. They finally pointed out that a full and nuanced understanding of data-intensive science can only be obtained by starting with the in-situ work and labor of scientific practice in all its manifold forms. Their work underscores a critical perspective: that the comprehensive grasp of data-intensive science is not merely about acknowledging the presence of invisible labor, but also about deciphering its unique characteristics when compared to other scientific endeavours.

Moreover, the importance of invisible labor in data-intensive science cannot be overstated. It is the often-uncredited work of data cleaning, metadata creation, and algorithm development that forms the backbone of robust scientific inquiry (Resnik et al., 2017; Shamoo, 2013). These tasks, while critical, are frequently overshadowed by the final published research outputs, which tend to be the metrics by which scientific success is commonly judged (Dance, 2012). The labor behind data maintenance, ensuring the integrity and accessibility of datasets, is equally vital, yet it remains largely invisible to those outside the immediate circle of data-intensive scientific practice.

In contemporary scientific research, data has become the central element propelling innovation and discovery. With the rapid evolution of big data technologies, data-intensive science has emerged as a prominent field, bringing revolutionary changes to various academic disciplines. This research methodology relies on the use of extensive datasets to drive experimentation, simulations, and analyses, simpler models with a lot of data supposedly trump more elaborate models with less data (Halevy et al., 2009), thereby enabling the acquisition of profound insights and knowledge. However, there is a paucity of research on data-intensive scientific activities. Approaching the understanding of data-intensive scientific activities from the perspective of division of labor and collaboration can better reveal the patterns of the fourth paradigm.

Evolution and adoption of CRediT

As the international community places increasing emphasis on the construction of research integrity, there is a growing call in the journal industry for the use of CRediT (Contribution Roles Taxonomy) to facilitate more transparent and granular descriptions of author contributions (Das & Das, 2020; Rahman & Verhagen, 2023; Udey, 2018). However, the forms and content of scientific research activities are evolving rapidly, the current taxonomy may need to evolve as science and the types of contributions that may become less or more important change (Allen et al., 2019). Therefore, for CORTs (Contributor Role Taxonomies), maintaining an up-to-date list of roles is essential to meet the evolving needs of users and is one of the factors that promote the adoption of CROTs (Vasilevsky et al., 2021). As emphasized by its developers, CRediT was initially designed as a contributor role taxonomy for life and physical sciences, and thus may not be suitable for all

disciplines(Allen et al., 2019), more specific roles should be added to the CRediT, and a more granular lexicon of contribution elements should be established. (Steele et al., 2021)。

Holcombe compared the author roles reflected by CRediT with the authorship criteria provided by the ICMJE (International Committee of Medical Journal Editors), elucidating the significance of the CRediT role taxonomy (Holcombe, 2019). Some scholars have discussed the boundaries of CRediT usage and proposed modifications. For instance, Alpi and Akers noted that CRediT does not currently describe all the roles played by librarians (Alpi & Akers, 2021). Larivière et al. employed the CRediT, combined with data from PLOS journals, to study the division of scientific labor from aspects such as author order, gender, and contribution combinations, highlighting the need for increased attention to labor division across different disciplines and research teams (Lariviere et al., 2021). Ding et al. proposed a new method for co-author credit allocation based on CRediT, demonstrating through empirical analysis that this approach can effectively prevent credit inflation and reasonably reflect author contributions, particularly mitigating the impact of the number of co-authors on the first author's credit (Ding et al., 2021).

Numerous scholars have taken CRediT as a foundation and, in conjunction with domain research or specific scenarios, have empirically conducted certain additions, deletions, and modifications to enhance its applicability. Some studies have pointed out the deficiencies of CRediT in the field of Randomized Controlled Trials (RCTs), with some attributing these to improper use by authors, such as not distinguishing between manuscript editing and drafting the initial manuscript (Steele et al., 2021). Others have identified inherent design flaws in CRediT, leading to the proposal of CRediT-RCT for researchers in this field (Zhang et al., 2019). Matarese and Shashok argued that CRediT overlooks some non-author contributions, including technical support, translation, and manuscript editing (Matarese & Shashok, 2019). In response, they proposed specific recommendations to improve three roles within CRediT (Investigation; Writing - Original Draft; Writing - Review & Editing) and to add two new roles (Technical support; Translating or editing the manuscript, as non-author). Alliez et al. emphasized the importance of software development, proposing nine more refined categories to better represent the tasks involved (design, debugging, maintenance, coding, architecture, documentation, testing, support, and management) (Alliez et al., 2020). Fitzgerald et al. summarized the shortcomings of the CRediT in the social sciences and proposed improvements, suggesting 12 librarian author role classifications (literature synthesis, conceptualization, methodology, instruments, software, investigation, data curation, data analysis, interpretation, visualization, writing, editing) (Fitzgerald et al., 2020).

In summary, existing research provides empirical analyses of CRediT across various disciplines and application scenarios, along with recommendations for improvement. The data paper, as an emerging form of academic communication, represents a unique type of scholarly output generated from data-intensive research activities, and the description and attribution of author contributions within this context also warrant attention.

Dataset and methods

Data source

We selected data papers published in the Data in Brief journal as our sample data source. The reason for choosing this particular data source is that Data in Brief is a purely data-focused interdisciplinary journal where all papers published are descriptions of datasets and serve as records of data-intensive scientific research activities. Additionally, Data in Brief utilizes CRediT to describe authors' contributions, which provides us with a consistent data source for analysis.

We utilized the web scraping tool Web Scraper (Scraper) to crawl and collect information related to data papers of *Data in Brief*. The collected data included the disciplinary fields of the articles, titles, DOIs, author names, author order, corresponding authors, and CRediT Author Statements, from 10 volumes (V41-V50) of *Data in Brief* published between 2022 and 2023, totalling 1,724

articles. After excluding corrigendum, articles without author contribution statements, and those not fully utilizing the CRediT format, we ultimately included 1,513 data papers for analysis, which comprised 7,697 CRediT contribution entries.

Research route

Scholars have conducted extensive research on author task division using CRediT (Lariviere et al., 2021; Zhang et al., 2019). Building upon their research approaches, we have formulated the research route for this paper, as detailed below:

（1）We initially present some basic information about the dataset, then employ Kendall's coefficient to examine the correlation between CRediT roles and conduct a statistical analysis of the symmetry in the roles undertaken by authors.

（2）Then, we investigate the general division of labor characteristics of data-intensive scientific research activities from the perspective of the CRediT roles involved in data papers. Firstly, the usage of CRediT roles, includes the number of CRediT roles used in data papers and the frequency of use for each role. Second, we examined the variation in CRediT roles with the number of authors per article, ranging from single-authored works to collaborative papers with up to 16 authors. Third, we designed the CRediT Concentration Index (CCI) and utilized the Gini coefficient to investigate the preference for CRediT roles across different disciplinary fields. The formula for calculating CCI is as follows:

$$CCI = \frac{U_{max_role}}{\sum U_{role}}$$

U_{max_role} represents the number of times the most frequently used role appears in a particular academic discipline, and U_{role} represents the usage count of each role within that discipline. It should be noted that to prevent excessive bias in the data, the calculation of CCI excludes the relevant data for the two writing roles, '*Writing – original draft*' and '*Writing – review and editing*'.

（3）Finally, we examine the detailed division of labor characteristics of data-intensive scientific research activities by considering the CRediT roles undertaken by key authors. Descriptive statistics were conducted based on whether an author was the corresponding author and whether they were the first author (without considering joint first authorship), and the chi-square test was used to examine if the differences between groups were significant. The number of roles undertaken by each author was treated as skewed data, and the Wilcoxon rank-sum test was employed to assess the significance of the differences between groups. All 14 CRediT roles were coded as binary variables. A logistic regression model was employed to examine the influence of individual roles on the status of the corresponding author and the first author. Due to the right-skewed distribution of the data, a generalized linear model with a Gamma distribution and a log link function was constructed to assess the impact of individual roles on the author order status.

Results

Overview

Demography of the dataset

As shown in Table 1, the dataset of data papers utilized in this study spans 24 disciplines, including Agricultural Sciences and Arts and Humanities, as designated by Data in Brief. The number of data papers varies across each discipline, with the highest count in Biological Sciences, totalling 205 papers, and the lowest in Mathematics, with only 3 papers. There is also a significant variation in the average number of authors per paper, ranging from 1.5 to 7.3 authors per paper. However, the average number of CRediT roles per paper and per author across different disciplines is relatively

stable, with most papers in the dataset involving an average of 6 to 10 CRediT roles per paper, and each author undertaking 3 to 5 CRediT roles.

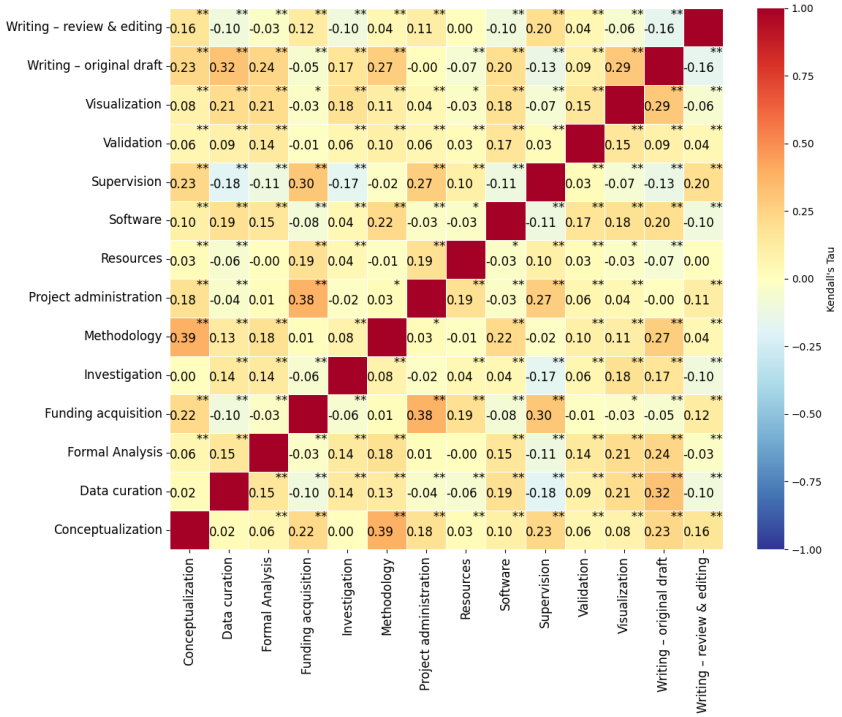
Disciplines	No. papers	Mean No. authors per paper	Mean No. roles per paper	Mean No. roles per author
Agricultural Sciences	164	5.0	8.5	3.3
Arts and Humanities	17	3.5	7.9	4.0
Biological Sciences	205	5.6	9.1	3.4
Business, Management, and Decision Sciences	64	3.3	8.5	4.1
Chemistry	58	5.2	8.9	3.1
Computer Science	166	4.5	9.3	3.7
Data Science	105	4.5	9.1	3.8
Earth and Planetary Sciences	90	6.0	9.2	3.5
Economics, Econometrics and Finance	34	2.9	8.7	4.9
Energy	43	5.6	9.2	3.4
Engineering	108	4.5	9.0	4.0
Environmental Science	100	5.5	8.8	3.3
Food Science	10	6.4	10.7	3.7
Health and Medical Sciences	103	7.3	9.2	3.4
Materials Science	56	4.6	9.2	3.5
Mathematics	3	2.3	11.7	7.3
Microbiology	25	5.8	8.9	3.1
Neuroscience	20	6.4	9.8	3.2
Omics	34	5.1	8.6	3.3
Pharmaceutical Sciences	6	7.2	7.5	2.4
Physical sciences	4	1.5	6.3	4.5
Plant Science	10	5.8	8.9	2.9
Psychology	30	4.3	8.1	3.8
Social Sciences	57	4.4	8.0	3.5

Table 1. Characteristics of CRediT roles from data papers in different disciplines

Correlation between CRediT roles

Figure 1 illustrates the correlations between various CRediT roles. Notably, the strongest correlation is observed between '*methodology*' and '*conceptualization*' (correlation coefficient = 0.39), followed by the correlation between '*funding acquisition*' and '*project administration*' (0.38), and the correlation between '*data curation*' and '*writing – original draft*' (0.32).

Figure 2 reflects the asymmetry in the associations between CRediT roles. More specifically, it shows the percentage of authors who have performed contribution A who have also performed contribution B (Lariviere et al., 2021). For instance, the figure shows that while 68.18% of authors who contributed to funding acquisition also reviewed and edited the manuscript, only 13.87% of authors who reviewed and edited the manuscript were involved in funding acquisition, representing the most asymmetric relationship. Following this is the asymmetry between theoretical and practical work, such as the observation that a significant portion of authors who undertook hands-on roles like '*formal analysis*' '*software*' '*validation*' and '*visualization*' also took on theoretical roles such as '*conceptualization*' and '*methodology*'. However, among the group of authors engaged in theoretical work, a relatively small fraction participated in the aforementioned practical tasks.



**. At the 0.01 level (two-tailed), the correlation is significant.

*. At the 0.05 level (two-tailed), the correlation is significant.

Figure 1. Heatmap showing the pairwise correlation of author roles defined in the CRediT

Contribution A \ Contribution B	Project administration	Funding acquisition	Formal Analysis	Resources	Software	Validation	Visualization	Investigation	Supervision	Writing –original draft	Data curation	Methodology	Conceptualization	Writing –review & editing
Project administration	/	8.27%	6.38%	30.44%	31.75%	11.54%	31.91%	13.34%	51.15%	49.35%	38.13%	48.36%	63.26%	40.75%
Funding acquisition	12.41%	/	48.03%	20.15%	23.71%	32.06%	19.16%	58.48%	25.43%	28.38%	31.94%	62.53%	42.87%	68.18%
Formal Analysis	8.28%	41.51%	/	12.74%	15.71%	31.32%	14.65%	58.81%	18.90%	22.82%	26.65%	66.77%	40.45%	67.62%
Resources	31.03%	13.68%	10.01%	/	30.53%	11.59%	37.53%	17.26%	54.21%	59.55%	54.38%	46.29%	51.21%	44.70%
Software	29.33%	14.59%	11.19%	27.66%	/	16.10%	30.31%	27.59%	39.83%	37.26%	40.74%	44.60%	49.28%	56.46%
Validation	13.48%	24.95%	28.20%	13.29%	20.36%	/	17.88%	35.09%	23.90%	20.46%	38.53%	41.87%	38.15%	52.58%
Visualization	27.80%	11.12%	9.84%	32.07%	28.58%	13.33%	/	14.75%	45.62%	52.10%	48.25%	44.33%	57.73%	48.68%
Investigation	8.59%	25.08%	29.19%	10.91%	19.23%	19.34%	10.91%	/	16.33%	18.76%	20.50%	57.17%	37.41%	69.55%
Supervision	26.04%	8.63%	7.42%	27.08%	21.96%	10.42%	26.67%	12.92%	/	50.17%	43.96%	39.21%	48.63%	44.29%
Writing –original draft	27.21%	10.42%	9.70%	32.22%	22.25%	9.66%	32.99%	16.06%	54.33%	/	47.02%	55.28%	59.52%	39.62%
Data curation	17.70%	9.87%	9.53%	24.76%	20.47%	15.31%	25.71%	14.77%	40.07%	39.57%	/	37.87%	44.44%	44.93%
Methodology	20.34%	17.52%	21.64%	19.10%	20.30%	15.07%	21.40%	37.34%	32.38%	42.15%	34.31%	/	63.25%	62.08%
Conceptualization	25.82%	11.66%	12.73%	20.51%	21.78%	13.33%	27.05%	23.71%	38.98%	44.05%	39.08%	61.39%	/	54.34%
Writing –review & editing	12.45%	13.87%	15.92%	13.40%	18.67%	13.75%	17.07%	32.99%	26.57%	21.94%	29.57%	45.09%	40.66%	/

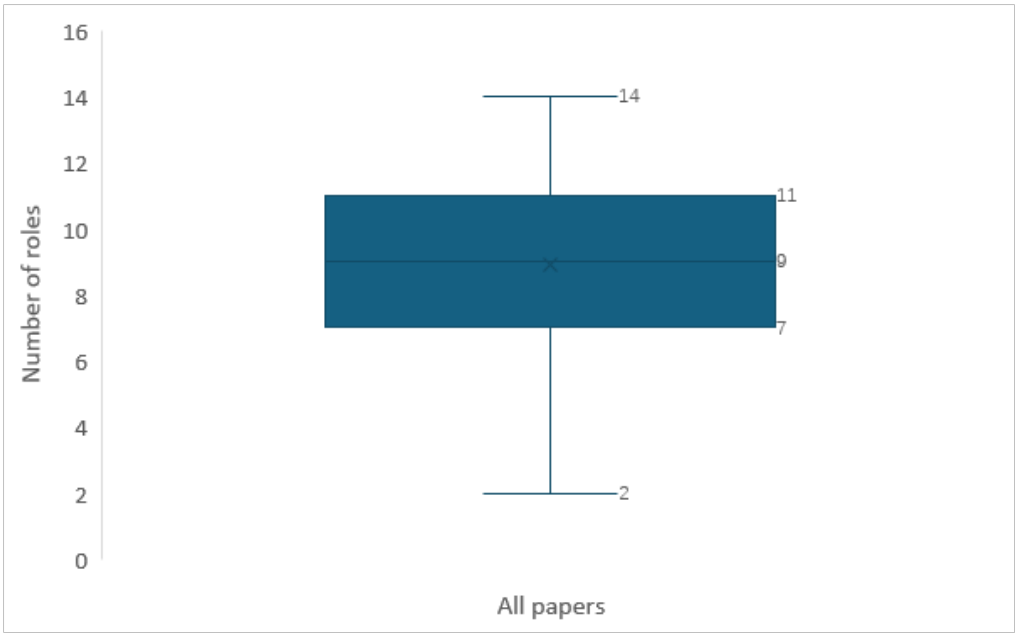
Figure 2. Percentage of authors who have performed contribution A who also have performed contribution B

CRediT roles involved in data papers

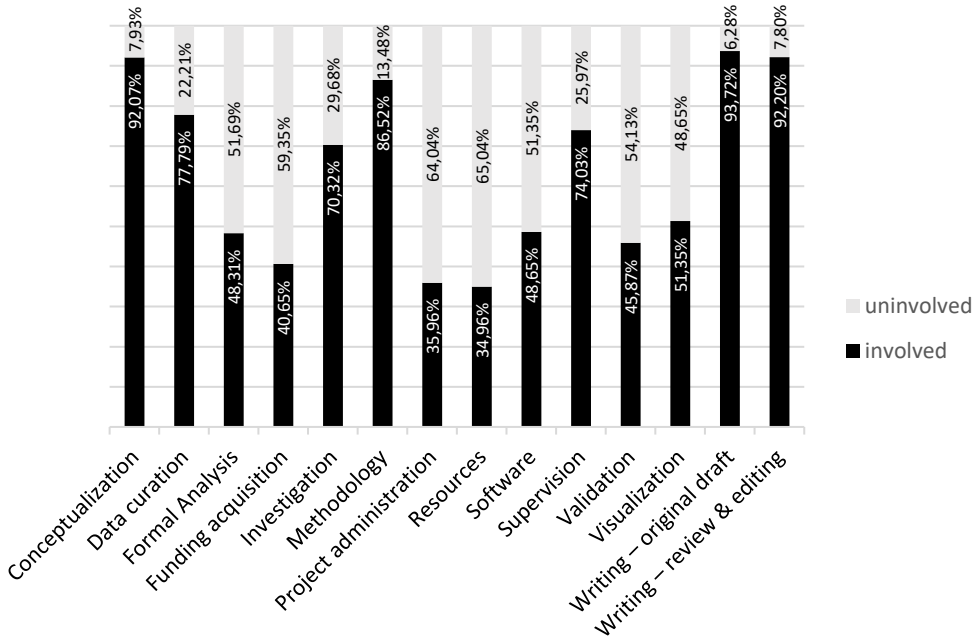
Usage of CRediT roles

Figure 3(a) presents the distribution of the number of CRediT roles involved across all articles. It can be observed that the majority of articles utilize approximately 9 CRediT roles to describe the research work. Figure 3(b) reflects the percentage of papers employing a particular CRediT role out of all data papers. It is evident that the usage of various roles in data papers is not uniform, the most frequently used roles are '*writing – original draft*', '*writing – review & editing*', and '*conceptualization*', all with a prevalence of over 90%. Following these are '*methodology*', which is associated with 86.52% of the papers, and roles such as '*data curation*', '*supervision*', and

'*investigation*', which are included in 77.79%, 74.03%, and 70.32% of the data papers, respectively. It is also noticeable that '*project administration*' and '*resources*' are less commonly utilized in data papers.



（a）



（b）

Figure 3. Statistical Overview of CRediT Role Usage

The division of labor varies with the number of authors

The division of labor in scientific research varies with the number of authors involved. Figure 4 presents the percentage of authors who have performed a given task, for papers between 1 and 16

authors (N=7296 papers, 94.8% of the data set). We assume that the sole author of a paper undertakes all 14 CRediT roles, and as the number of authors increases, the distribution of different tasks among authors begins to shift. The changes can be broadly categorized into three types: one type of role is consistently carried out by a smaller proportion of the team (dashed line), such as '*project administration*', '*funding acquisition*', and '*resources*', which are consistently undertaken by about one-fifth of the total team members; a second type of role is consistently undertaken by a larger proportion of team members (thick line), such as '*writing – review and editing*' and '*investigation*', which are consistently handled by about one-third to one-half of the authors; and a third type of role sees a significant decrease in the proportion of authors undertaking it as the number of authors increases (solid line), such as '*conceptualization*', '*methodology*', and '*writing – original draft*', which are initially undertaken by half or more of the authors when the number of authors is low, but drop to being handled by only about one-fifth of the authors when the team size exceeds 10.

It is noteworthy that as the size of the author team increases, particularly when it reaches more than 10 members, the proportion of authors assuming CRediT roles begins to fluctuate. However, roles associated with data collection or analysis tend to emerge with disproportionately high levels of involvement, such as '*validation*' and '*formal analysis*'. Notably, the role of '*investigation*' stands out; when the number of authors exceeds 11, the proportion of authors undertaking the '*investigation*' role surpasses that of roles like '*conceptualization*', '*methodology*', and '*data curation*'. Alongside '*writing – review and editing*', '*investigation*' becomes one of the roles with the highest proportion of authors involved.

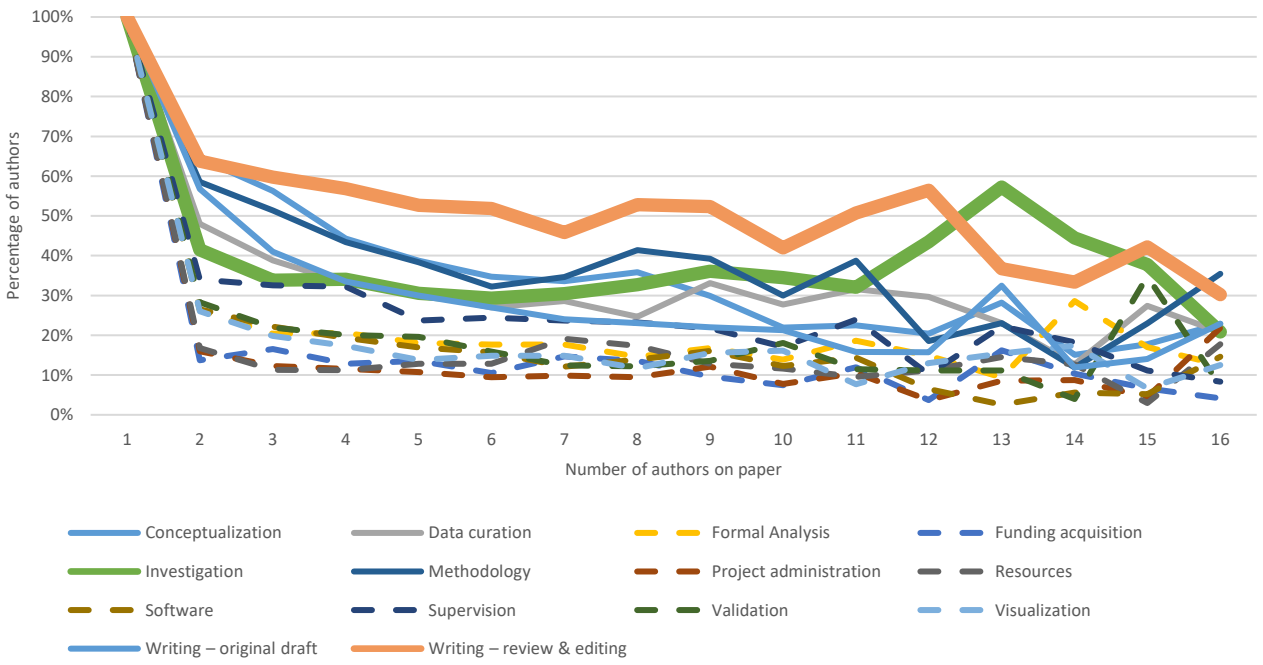


Figure 4. Statistical overview of CRediT role variation with the number of authors

CRediT role concentration across different disciplines

Disciplinary differences exist in the concentration of CRediT roles, and Figure 5 reflects the concentration of CRediT roles involved in articles from the top 13 most represented fields in the sample dataset (N=6767, 87.9%).

Overall, the concentration of CRediT roles in data papers across various academic disciplines fluctuates between 10% and 20%. The highest concentration is observed in the social sciences

field, at 19.18%, while the lowest is in the computer science field, at 13.44%. Relatively speaking, social science fields, such as social sciences, business, management, and decision sciences, exhibit significantly higher CRediT role concentration compared to some natural science fields, like biological sciences and energy. The Gini coefficient, a measure of inequality, further highlights the disparities in CRediT role distribution across disciplines. Higher values of the Gini coefficient indicate greater inequality in role distribution, which aligns with the observed higher concentration in social science fields (0.31). This metric underscores the need for a more balanced distribution of CRediT roles across all academic disciplines.

Furthermore, the specific CRediT roles that authors concentrate on vary by discipline. Upon closer examination, with the exception of the '*writing – review and editing*' role, authors in social science fields predominantly focus on the '*conceptualization*' role, whereas authors in natural science fields are more inclined toward the '*investigation*' and '*methodology*' roles (Figure 6).

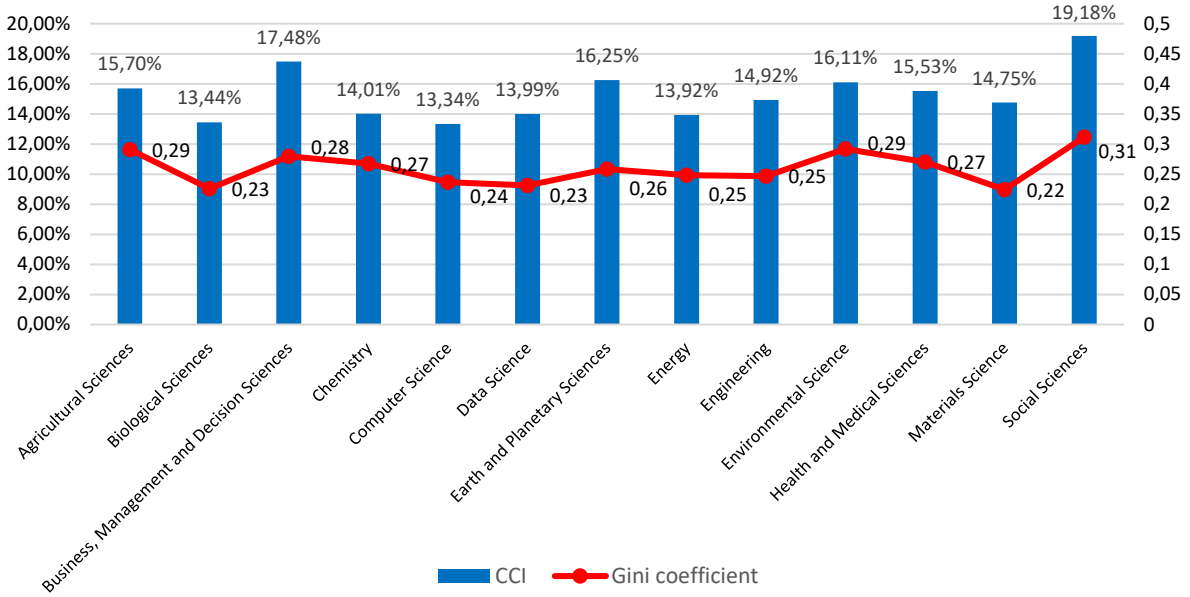


Figure 5. CRediT Concentration Index (CCI) and Gini coefficient across different disciplines

CRediT \ Discipline	Agricultural Sciences	Biological Sciences	Business, Management and Decision Sciences	Chemistry	Computer Science	Data Science	Earth and Planetary Sciences	Energy	Engineering	Environmental Science	Health and Medical Sciences	Materials Science	Social Sciences
Conceptualization	37.20%	33.20%	55.40%	33.30%	38.60%	38.90%	34.30%	34.20%	41.00%	40.10%	33.50%	39.20%	50.00%
Data curation	29.80%	30.60%	39.90%	27.40%	35.50%	35.70%	33.30%	25.80%	29.20%	27.10%	32.90%	23.10%	31.00%
Formal Analysis	16.80%	21.60%	21.60%	13.50%	14.60%	12.40%	24.40%	14.20%	17.80%	17.30%	20.40%	18.40%	15.50%
Funding acquisition	11.00%	16.60%	12.70%	10.60%	10.90%	12.40%	12.70%	12.90%	15.50%	12.60%	6.00%	13.70%	12.30%
Investigation	34.90%	35.00%	25.40%	33.70%	29.10%	28.80%	34.80%	32.90%	45.30%	34.20%	40.80%	39.20%	27.40%
Methodology	39.10%	33.30%	45.50%	33.30%	37.80%	41.00%	44.50%	36.20%	47.00%	37.90%	33.60%	29.40%	44.00%
Project administration	7.80%	10.70%	10.80%	9.20%	12.60%	14.10%	10.80%	8.30%	11.80%	7.70%	11.90%	12.90%	7.10%
Resources	9.40%	14.50%	12.20%	13.20%	12.70%	12.80%	17.10%	9.20%	14.50%	14.90%	14.80%	16.50%	12.70%
Software	13.10%	12.00%	24.40%	7.90%	26.60%	24.80%	11.70%	19.60%	19.90%	9.50%	13.50%	13.30%	13.90%
Supervision	22.10%	23.80%	31.00%	26.10%	31.60%	27.60%	15.10%	25.40%	29.60%	22.00%	24.90%	26.30%	19.00%
Validation	16.50%	14.10%	19.70%	17.80%	24.80%	25.40%	19.60%	27.10%	22.60%	11.50%	13.20%	13.70%	15.10%
Visualization	13.10%	14.70%	18.30%	14.20%	14.60%	19.20%	15.60%	14.60%	20.70%	14.00%	17.20%	20.00%	12.70%
Writing -original draft	29.10%	26.10%	40.40%	26.40%	32.30%	31.80%	25.90%	25.40%	33.50%	25.50%	28.30%	25.50%	35.30%
Writing -review & editing	49.00%	51.90%	55.90%	45.20%	46.20%	54.90%	46.90%	50.40%	53.60%	56.30%	55.70%	54.90%	54.80%

Figure 6. Percentage of authors per CRediT Role across different disciplines

CRediT roles undertaken by key authors

Factors associated with the corresponding author

A statistical analysis of 7967 author contributions across 1513 articles revealed that (Table 2), in general, apart from writing roles, authors were most involved in '*methodology*' (38.9%) and '*conceptualization*' (37.8%), while a smaller proportion of authors participated in "Project administration" (10.6%). The median number of roles per author was 3 (IQR: 2-5).

For corresponding authors, the proportion of those assuming the '*conceptualization*' role is significantly higher compared to other roles, accounting for 65.0%, and they are more likely to take on any of the 14 CRediT roles relative to non-corresponding authors. It is noteworthy that the majority of corresponding authors are also first authors, with 899 out of 1958 (45.9%). Corresponding authors take on a greater number of roles than non-corresponding authors [5 (3-6) vs. 3 (2-4); P<0.001], and they are ranked higher in author order [2 (1-4) vs. 4 (2-6); P < 0.001]. Additionally, there is no significant difference in involvement in the '*investigation*' and '*resources*' roles based on whether an author is a corresponding author or not (P>0.1).

Variables	Overall（n=7697）	Non-corresponding author (n=5739)	Corresponding author (n=1958)	P
Conceptualization, n (%)	2906 (37.8)	1633 (28.5)	1273 (65.0)	<0.001
Data_curation, n (%)	2400 (31.2)	1623 (28.3)	777 (39.7)	<0.001
Formal_Analysis, n (%)	1403 (18.2)	917 (16.0)	486 (24.8)	<0.001
Funding_acquisition, n (%)	942 (12.2)	531 (9.3)	411 (21.0)	<0.001
Investigation, n (%)	2633 (34.2)	1937 (33.8)	696 (35.5)	0.156
Methodology, n (%)	2994 (38.9)	1932 (33.7)	1062 (54.2)	<0.001
Project_administration, n (%)	814 (10.6)	479 (8.3)	335 (17.1)	<0.001
Resources, n (%)	1046 (13.6)	762 (13.3)	284 (14.5)	0.184
Software, n (%)	1222 (15.9)	799 (13.9)	423 (21.6)	<0.001
Supervision, n (%)	1898 (24.7)	1199 (20.9)	699 (35.7)	<0.001
Validation, n (%)	1323 (17.2)	909 (15.8)	414 (21.1)	<0.001
Visualization, n (%)	1199 (15.6)	720 (12.5)	479 (24.5)	<0.001
Writing-original draft, n (%)	2216 (28.8)	1174 (20.5)	1042 (53.2)	<0.001
Writing-review & editing, n (%)	4001 (52.0)	2886 (50.3)	1115 (56.9)	<0.001
First_author, n (%)	1513 (19.7)	614 (10.7)	899 (45.9)	<0.001
roles_number, median (IQR)	3 (2, 5)	3 (2, 4)	5 (3, 6)	<0.001
Author_order, median (IQR)	3 (2, 5)	4 (2, 6)	2 (1, 4)	<0.001

IQR, interquartile range.

Table 2. Comparison of CRediT roles between corresponding author and other co-authors

The logistic regression model indicates that authors who perform '*conceptualization*' are twice as likely to be corresponding authors (OR: 2.33; 95% CI: 2.04-2.67; P<0.001). Furthermore, authors engaged in '*funding acquisition*' (OR: 1.91; 95% CI: 1.60-2.29; P<0.001), '*supervision*' (OR: 2.26; 95% CI: 1.95-2.61; P<0.001), '*writing - original draft*' (OR: 2.37; 95% CI: 2.03-2.77; P<0.001), and those

serving as first authors (OR: 4.65; 95% CI: 3.94-5.50; P<0.001) are more likely to be corresponding authors (Table 3).

	OR (95% CI)	P
Conceptualization	2.33 (2.04, 2.67)	<0.001
Data_curation	1.12 (0.97, 1.29)	0.115
Formal_Analysis	1.00 (0.85, 1.17)	1
Funding_acquisition	1.91 (1.60, 2.29)	<0.001
Investigation	0.75 (0.66, 0.86)	<0.001
Methodology	0.95 (0.83, 1.09)	0.461
Project_administration	1.19 (0.98, 1.45)	0.078
Resources	1.05 (0.87, 1.26)	0.599
Software	1.04 (0.87, 1.23)	0.68
Supervision	2.26 (1.95, 2.61)	<0.001
Validation	1.02 (0.87, 1.20)	0.819
Visualization	1.23 (1.03, 1.45)	0.018
Writing-original draft, n (%)	2.37 (2.03, 2.77)	<0.001
Writing-review & editing, n (%)	1.37 (1.21, 1.56)	<0.001
First_author	4.65 (3.94, 5.50)	<0.001

OR odds ratio; CI, confidence interval.

Table 3. Logistic regression model investigating the factors associated with the corresponding author role

Factors associated with the first author

Due to a significant overlap between the first authors and corresponding authors, the roles undertaken by the first authors are similar to those of the corresponding authors. However, there are subtle differences. There are no significant differences between first authors and non-first authors in the roles of '*funding acquisition*' and '*project administration*' (Table 4).

Variables	Overall（n=7697）	Non-first author (n=6184)	First author (n=1513)	P
Conceptualization, n (%)	2906 (37.8)	1910 (30.9)	996 (65.8)	<0.001
Data_curation, n (%)	2400 (31.2)	1541 (24.9)	859 (56.8)	<0.001
Formal_Analysis, n (%)	1403 (18.2)	839 (13.6)	564 (37.3)	<0.001
Funding_acquisition, n (%)	942 (12.2)	795 (12.9)	147 (9.7)	0.001
Investigation, n (%)	2633 (34.2)	1863 (30.1)	770 (50.9)	<0.001
Methodology, n (%)	2994 (38.9)	1949 (31.5)	1045 (69.1)	<0.001
Project_administration, n (%)	814 (10.6)	660 (10.7)	154 (10.2)	0.607
Resources, n (%)	1046 (13.6)	913 (14.8)	133 (8.8)	<0.001
Software, n (%)	1222 (15.9)	728 (11.8)	494 (32.7)	<0.001
Supervision, n (%)	1898 (24.7)	1686 (27.3)	212 (14.0)	<0.001
Validation, n (%)	1323 (17.2)	968 (15.7)	355 (23.5)	<0.001
Visualization, n (%)	1199 (15.6)	658 (10.6)	541 (35.8)	<0.001
Writing-original draft, n (%)	2216 (28.8)	980 (15.8)	1236 (81.7)	<0.001
Writing-review & editing, n (%)	4001 (52.0)	3379 (54.6)	622 (41.1)	<0.001
Author_corresponding, n (%)	1958 (25.4)	1059 (17.1)	899 (59.4)	<0.001
roles_number, median (IQR)	3 (2, 5)	3 (2, 4)	5 (4, 7)	<0.001

IQR, interquartile range.

Table 4. Comparison of CRediT roles between the first author and other co-authors

The logistic regression model indicates a significant correlation between the role of '*writing - original draft*' and being designated as the first author (OR: 11.23; 95% CI: 9.56-13.23; P<0.001). However, first authors are seldom involved in '*writing - review and editing*' (OR: 0.71; 95% CI: 0.61-0.84; P<0.001), '*supervision*' (OR: 0.53; 95% CI: 0.42-0.66; P<0.001), or '*resources*' (OR: 0.57; 95% CI: 0.43-0.74; P<0.001) (Table 5).

	OR (95% CI)	P
Conceptualization	3.50 (2.94, 4.17)	<0.001
Data_curation	1.60 (1.36, 1.87)	<0.001
Formal_Analysis	1.72 (1.44, 2.05)	<0.001
Funding_acquisition	0.86 (0.65, 1.14)	0.304
Investigation	1.49 (1.27, 1.74)	<0.001
Methodology	1.88 (1.59, 2.22)	<0.001
Project_administration	0.89 (0.66, 1.19)	0.423
Resources	0.57 (0.43, 0.74)	<0.001
Software	1.60 (1.33, 1.93)	<0.001
Supervision	0.53 (0.42, 0.66)	<0.001
Validation	0.87 (0.71, 1.06)	0.179
Visualization	1.79 (1.48, 2.15)	<0.001
Writing-original draft, n (%)	11.23 (9.56, 13.23)	<0.001
Writing-review & editing, n (%)	0.71 (0.61, 0.84)	<0.001

OR odds ratio; CI, confidence interval.

Table 5. Logistic regression model investigating the factors associated with the first author designation

Factors associated with the author order

Figure 7 revealed a significant correlation between the position of each author in an article and the number of roles they undertake (R^2=0.073, P<0.001). The results of the generalized linear model analysis (Table 6) indicate a significant correlation between '*conceptualization*', '*data curation*', '*software*', and '*visualization*' with the order of authorship. For instance, authors who contributed to the '*software*' are likely to be ranked higher compared to those who did not participate in this role, with a coefficient of 0.28 (95% CI: 0.22-0.34).

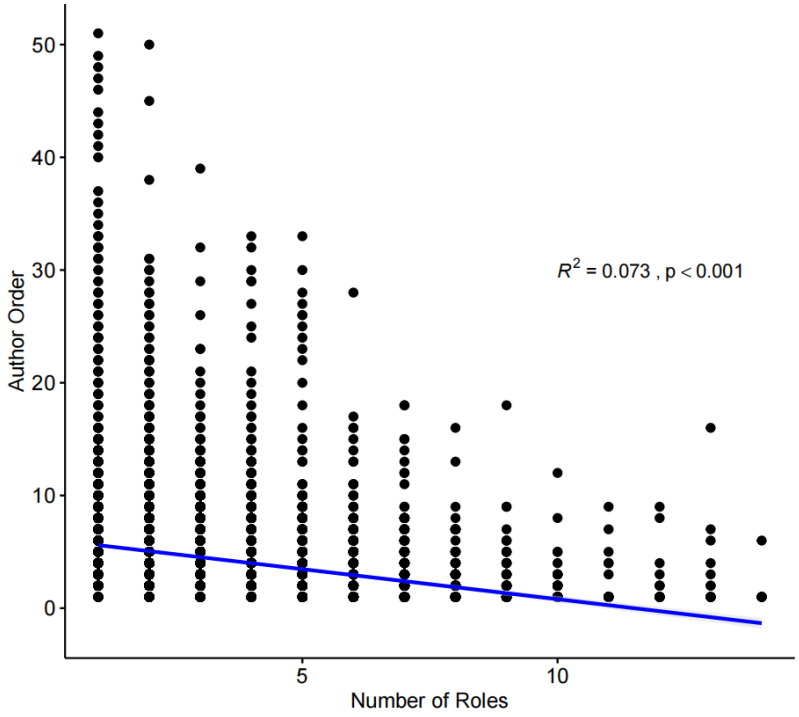


Figure 7. Scatter plot showing the correlation between the author order and the number of roles per author. There was a significant correlation (R^2 =0.073; P<0.001) between the two variables.

	Coefficient (95% CI)	P
Conceptualization	-0.23 (-0.28, -0.18)	<0.001
Data_curation	-0.15 (-0.19, -0.10)	<0.001
Formal_Analysis	-0.10 (-0.16, -0.04)	<0.05
Funding_acquisition	0.07 (-0.00, 0.14)	0.059
Investigation	-0.07 (-0.11, -0.03)	<0.05
Methodology	-0.08 (-0.12, -0.03)	<0.05
Project_administration	0.04 (-0.03, 0.12)	0.26
Resources	0.09 (0.03, 0.15)	<0.05
Software	**-0.28 (-0.34, -0.22)**	**<0.001**
Supervision	0.04 (-0.01, 0.09)	0.142
Validation	-0.04 (-0.09, 0.02)	0.193
Visualization	-0.16 (-0.22, -0.10)	<0.001
Writing-original draft	-0.59 (-0.64, -0.54)	<0.001
Writing-review & editing	-0.10 (-0.15, -0.06)	<0.001

Table 6. Generalized linear model investigating the factors associated with the author order

Discussion

This study employs CRediT to explore the division of labor in data-intensive scientific research activities, represented by data papers. According to our result, the pattern of labor division in data articles is different from traditional research papers in many ways. For instance, in data papers, the core contribution of '*conceptualization*' is most strongly correlated with the contribution of '*methodology*', while it is '*funding acquisition*' with research articles in the RCT field (Zhang et al., 2019). However, the roles that play a significant part in data paper collaborations show no significant difference from those in research articles: almost all articles involve roles such as '*conceptualization*', '*writing – original draft*', and '*writing – review and editing*' (Lariviere et al., 2021). Furthermore, some roles exhibit high autocorrelation and lack distinct differentiation, suggesting that they could be considered for consolidation within the context of contributor roles in data papers.

Data-intensive science is a collaborative endeavour, and our research indicates that as the number of co-authors increases, there is a corresponding rise in the number of individuals taking on roles such as investigation, validation, and formal analysis. In contrast, the number of individuals assuming core roles like conceptualization, methodology, and Writing – original draft, or managerial roles such as project administration, funding acquisition, and resources, remains relatively stable. This suggests that the primary driver of team size expansion is the need for practical tasks such as data handling and investigation, while the number of scientific leaders remains scarce (Robinson-Garcia et al., 2020). As Larivière et al. have posited, '*the bureaucratization of science can be considered as an inevitable consequence of the ubiquity of collaborative science*' (Larivière et al., 2015).

The publication of data papers reflects the work patterns and research cultures across various disciplines. Different academic fields exhibit varying degrees of concentration in the use of CRediT roles, with data papers in the social sciences generally showing higher concentrations of CRediT role usage compared to those in the natural sciences. We suggest that the differences in research methodologies, types of data, and the nature of research questions are the primary causes of this phenomenon. For instance, data papers in the social sciences more prominently represent the '*paper*' dimension, with authors tending to focus on the '*conceptualization*' role, while data papers in the natural sciences emphasize the '*data*' dimension more, with authors concentrating more on '*investigation*' and '*methodology*'.

The attribution of academic credit is one of the key concerns for researchers, with a common understanding that corresponding authors and first authors make significant contributions and play major leadership roles (Bhandari et al., 2014; Perneger et al., 2017; Teixeira da Silva, 2021; Yang et al., 2017). The statistical analysis indicates that in data papers, corresponding authors often undertake more leadership and coordination tasks and are more likely to be involved in key roles such as '*conceptualization*', '*funding acquisition*', and '*supervision*'. These roles pertain not only to the initial design and theoretical construction of the research but also encompass the supervision of the research process and financial support. The first authors, on the other hand, exhibit higher engagement in roles like '*writing - original draft*' and '*data curation*', which are directly related to the implementation of the research and the accuracy of the data. Comparatively, corresponding authors are more likely than first authors to take on any given role, implying that they act as versatile players within the team (Lu et al., 2020) and are more likely to be the corresponding authors of a data paper. Furthermore, our research also reveals that the more roles an author takes on, and their involvement in key data processing roles such as '*data curation*', '*software*', and '*visualization*', the more it aids in the author's ranking in terms of by-line order.

Our research indicates that some technical contributions related to data processing are clearly very important for the attribution of credit to authors of data papers. However, in the CRediT, the connotations of roles such as '*data curation*' encompass both work requiring profound professional

background knowledge, such as data annotation, and technical tasks like data cleaning and maintenance. The roles of technical contributions (Smith, 2023) and human intervention are not adequately differentiated and described. This could lead to a situation where some authors' significant contributions are not recognized as they should be, while others enjoy excessive credit due to their role designation. This situation could be more pronounced in data papers, ultimately affecting the fairness of academic credit attribution for data papers.

CRediT, as a standardized method for describing author contributions, aids in clarifying the specific contributions of each author and determining the data responsibilities of different authors in data papers. However, the content standards and organizational forms of data papers differ from those of traditional research articles (Callaghan et al., 2012). Data papers focus more on the collection, processing, and presentation of data, some of which do not have direct corresponding categories within the 14 CRediT roles. The majority of data papers use only about 9 roles to describe the work conducted, indicating that CRediT has limitations in fully capturing the research workflow. Some roles, such as '*validation*' and '*software*', which ensure data quality and reuse value, are missing in most data papers. This leads to an underestimation or neglect of the actual significance and unique contributions of these roles in research work, affecting a comprehensive understanding of the entire research process. This uneven usage also implies that CRediT fails to accurately and consistently provide all important types of contributions when describing the research work in data papers (Alliez et al., 2020; Fitzgerald et al., 2020; Matarese & Shashok, 2019).

Conclusion

Over the last few decades, data-intensive scientific research has become increasingly prevalent, with some of its labor division characteristics reflected through the publication of data papers. Our study arrives at the following main conclusions:

(1) Data papers rarely make full use of the 14 CRediT roles to describe author contributions, with '*project administration*' and '*resources*' being unmentioned in over half of the data paper samples.

(2) Team size and discipline have a significant impact on the labor division of data-intensive scientific research activities. The need for data collection and analysis is the main reason for the expansion of team size, which is particularly evident in the natural sciences, where authors' roles are more concentrated on '*investigation*' and '*methodology*'.

(3) Corresponding authors and first authors continue to take on core roles, such as '*methodology*' and '*conceptualization*', but at the same time, undertaking data analysis and processing-related tasks, such as '*software*', helps authors advance in the author order of data papers.

Data papers provide an excellent window into studying data-intensive scientific research activities, and CRediT offers a useful framework for characterizing data-centric scientific workflows, but it requires refinement to reflect the characteristics of data papers and the diversity of research work more comprehensively. Developing a taxonomy of contributor roles specific to data papers—DP-CRediT (Data paper-CRediT)—could be a good option. For instance, it might be beneficial to add specific roles such as '*data collection*' and '*metadata management*' to accurately reflect the contributions of these key steps in data papers. Moreover, the importance of technical work such as software development and data processing is increasingly recognized. Alliez et al. argue that since software development in research involves '*significant innovation*', there is a need for appropriate human intervention and qualitative information to ensure accurate reporting of contributions (Alliez et al., 2020). Building on the existing roles of '*software*' and '*formal analysis*', it could be considered to further refine technical contributions by introducing roles such as '*algorithm development*' and '*data engineering*' to describe technical work more comprehensively. Adding roles like '*critical analysis*' or '*data interpretation*' could emphasize the role of human intervention in data analysis and interpretation, ensuring a balance between technical

contributions and human judgment. While refining roles, it is also necessary to maintain consistency among existing CRediT roles. Some CRediT roles describe project-level tasks while others describe paper-level tasks(Hosseini, Gordijn, et al., 2023), which could lead to an imbalance in the allocation of academic credit.

Acknowledgements

This work was funded by the project "Publishing Model linked Research Paper and Scientific Data and its application in FAIR-compliant Manner"（23BXW097）supported by National Social Science Foundation of China

About the authors

YANG Heng is a Ph.D. candidate at the National Science Library, Chinese Academy of Sciences. His research interests include scientific data management, data publishing and dissemination, FAIR principles, and related areas. He can be contacted at yangheng@mail.las.ac.cn

YU Yonglin is a master's candidate at the National Science Library, Chinese Academy of Sciences. Her research interests include semantic publishing, scientific information editing and dissemination, and related fields. She can be contacted at yuyonglin@mail.las.ac.cn

LIU Fenghong is a research librarian at the National Science Library, Chinese Academy of Sciences. Her research interests include data publishing, scientific data management, FAIR principles, semantic publishing, and related areas. She can be contacted at liufh@mail.las.ac.cn

References

Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship [Article]. Learned Publishing, 32(1), 71-74. https://doi.org/10.1002/leap.1210

Alliez, P., Di Cosmo, R., Guedj, B., Girault, A., Hacid, M.-S., Legrand, A., & Rougier, N. (2020). Attributing and Referencing (Research) Software: Best Practices and Outlook From Inria [Article]. Computing in Science & Engineering, 22(1), 39-51. https://doi.org/10.1109/mcse.2019.2949413

Alpi, K. M., & Akers, K. G. (2021). CRediT for authors of articles published in the Journal of the Medical Library Association [Editorial Material]. Journal of the Medical Library Association, 109(3), 362-364. https://doi.org/10.5195/jmla.2021.1294

Benhamed, O. M., Burger, K., Kaliyaperumal, R., da Silva Santos, L. O. B., Suchánek, M., Slifka, J., & Wilkinson, M. D. (2023). The FAIR Data Point: Interfaces and Tooling. Data Intelligence, 5(1), 184-201. https://doi.org/10.1162/dint_a_00161

Bierer, B. E., Crosas, M., & Pierce, H. H. (2017). Data Authorship as an Incentive to Data Sharing. New England Journal of Medicine, 376(17), 1684-1687. https://doi.org/doi:10.1056/NEJMsb1616595

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P. J., Bowie, R. C., Leadbetter, A. M., Lowry, R. K., Moncoiffe, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. G. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. Int. J. Digit. Curation, 7, 107-113.

Cheruvelil, K. S., & Soranno, P. A. (2018). Data-Intensive Ecological Research Is Catalyzed by Open Science and Team Science. BioScience, 68(10), 813-822. https://doi.org/10.1093/biosci/biy097

Dance, A. (2012). Authorship: Who's on first? Nature, 489(7417), 591-593. https://doi.org/10.1038/nj7417-591a

Das, N., & Das, S. (2020). 'Author Contribution Details' and not 'Authorship Sequence' as a merit to determine credit: A need to relook at the current Indian practice [Review]. National Medical Journal of India, 33(1), 24-30, Article Pmid 33565483. https://doi.org/10.4103/0970-258x.308238

Ding, J., Liu, C., Zheng, Q., & Cai, W. (2021). A new method of co-author credit allocation based on contributor roles taxonomy: proof of concept and evaluation using papers published in PLOS ONE [Article]. Scientometrics, 126(9), 7561-7581. https://doi.org/10.1007/s11192-021-04075-x

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. Computer Supported Cooperative Work (CSCW), 19(3), 355-375. https://doi.org/10.1007/s10606-010-9117-8

Fitzgerald, S., Budd, J., Beile, P., & Kaspar, W. (2020). Modeling Transparency in Roles: Moving from Authorship to Contributorship. 2020(7). https://doi.org/10.5860/crl.81.7.1056

Greenberg, J., McClellan, S., Rauch, C., Zhao, X., Kelly, M., An, Y., Kunze, J., Orenstein, R., Porter, C., Meschke, V., & Toberer, E. (2023). Building Community Consensus for Scientific Metadata with YAMZ. Data Intelligence, 5(1), 242-260. https://doi.org/10.1162/dint_a_00211

Greenberg, J., Wu, M., Liu, W., & Liu, F. (2023). Metadata as Data Intelligence. Data Intelligence, 5(1), 1-5. https://doi.org/10.1162/dint_e_00212

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2), 8-12. https://doi.org/10.1109/MIS.2009.36

Holcombe, A. O. (2019). Contributorship, Not Authorship: Use CRediT to Indicate Who Did What [Article]. Publications, 7(3), Article 48. https://doi.org/10.3390/publications7030048

Hosseini, M., Colomb, J., Holcombe, A. O., Kern, B., Vasilevsky, N. A., & Holmes, K. L. (2023). Evolution and adoption of contributor role ontologies and taxonomies [Article]. Learned Publishing, 36(2), 275-284. https://doi.org/10.1002/leap.1496

Hosseini, M., Gordijn, B., Wafford, Q. E., & Holmes, K. L. L. (2023). A systematic scoping review of the ethics of Contributor Role Ontologies and Taxonomies [Review; Early Access]. Accountability in Research-Ethics Integrity and Policy. https://doi.org/10.1080/08989621.2022.2161049

ICSU. Priority Area Assessment on Scientific Data and Information. ICSU. https://council.science/publications/priority-area-assessment-on-scientific-data-and-information/

Kalager, M., Adami, H.-O., & Bretthauer, M. (2016). Recognizing Data Generation. New England Journal of Medicine, 374(19), 1898-1898. https://doi.org/doi:10.1056/NEJMc1603789

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., da Silva Santos, L. O. B., & Roos, M. (2020). The "A" of FAIR – As Open as Possible, as Closed as Necessary. Data Intelligence, 2(1-2), 47-55. https://doi.org/10.1162/dint_a_00027

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. Journal of the Association for Information Science and Technology, 66(7), 1323-1332. https://doi.org/10.1002/asi.23266

Lariviere, V., Pontille, D., & Sugimoto, C. R. (2021). Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT) [Article]. Quantitative Science Studies, 2(1), 111-128. https://doi.org/10.1162/qss_a_00097

Lenhardt, W. C., Conway, M., Scott, E., Blanton, B., Krishnamurthy, A., Hadzikadic, M., Vouk, M., Wilson, A., & Ieee. (2016, 2016

Sep 13-15). Cross-Institutional Research Cyberinfrastructure for Data Intensive Science.IEEE High Performance Extreme Computing Conference [2016 ieee high performance extreme computing conference (hpec)]. IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA.

Lo, B., & DeMets, D. L. (2016). Incentives for Clinical Trialists to Share Data. New England Journal of Medicine, 375(12), 1112-1115. https://doi.org/doi:10.1056/NEJMp1608351

Lu, C., Zhang, Y., Ahn, Y.-Y., Ding, Y., Zhang, C., & Ma, D. (2020). Co-contributorship Network and Division of Labor in Individual Scientific Collaborations [Article]. Journal of the Association for Information Science and Technology, 71(10), 1162-1178. https://doi.org/10.1002/asi.24321

Matarese, V., & Shashok, K. (2019). Transparent Attribution of Contributions to Research: Aligning Guidelines to Real-Life Practices [Article]. Publications, 7(2), Article 24. https://doi.org/10.3390/publications7020024

Nielsen, M. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Nature, 462(7274), 722-723. https://doi.org/10.1038/462722a

NISO. Contributor Roles Taxonomy. NISO. https://credit.niso.org/

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the Reuse of Scientific Data. Data Science Journal. https://doi.org/10.5334/dsj-2017-008

Pietsch, W. (2015). Aspects of Theory-Ladenness in Data-Intensive Science. Philosophy of Science, 82(5), 905-916. https://doi.org/10.1086/683328

Rahman, M. T., & Verhagen, J. V. (2023). Implementing Quantitative Declarations of Authorship Contribution: A Call to Action [Article]. Journal of Scientometric Research, 12(2), 431-435. https://doi.org/10.5530/jscires.12.2.039

Ramachandran, R., Rushing, J., Lin, A., Conover, H., Li, X., Graves, S., Nair, U. S., Kuo, K. S., & Smith, D. K. (2013). Data Prospecting–A Step Towards Data Intensive Science. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(3), 1233-1241. https://doi.org/10.1109/JSTARS.2013.2248133

Resnik, D. B., Elliott, K. C., Soranno, P. A., & Smith, E. M. (2017). Data-Intensive Science and Research Integrity. Accountability in Research-Ethics Integrity and Policy, 24(6), 344-358. https://doi.org/10.1080/08989621.2017.1327813

Robinson-Garcia, N., Costas, R., Sugimoto, C. R., Larivière, V., & Nane, G. F. (2020). Task specialization across research careers. Elife, 9, e60586. https://doi.org/ARTN e60586

10.7554/eLife.60586

Schultes, E., Roos, M., Bonino da Silva Santos, L. O., Guizzardi, G., Bouwman, J., Hankemeier, T., Baak, A., & Mons, B. (2022). FAIR Digital Twins for Data-Intensive Research. Front Big Data, 5, 883341. https://doi.org/10.3389/fdata.2022.883341

Scraper, W. About us. Web Scraper. https://www.webscraper.io/about-us

Scroggins, M. J., & Pasquetto, I. V. (2020). Labor Out of Place: On the Varieties and Valences of (In)visible Labor in Data-Intensive Science. Engaging Science Technology and Society, 6, 111-132. https://doi.org/10.17351/ests2020.341

Shamoo, A. E. (2013). Data Audit as a Way to Prevent/Contain Misconduct. Accountability in Research-Policies and Quality Assurance, 20(5-6), 369-379. https://doi.org/10.1080/08989621.2013.822259

Smith, E. (2023). "Technical" Contributors and Authorship Distribution in Health Science [Article]. Science and Engineering Ethics, 29(4), Article 22. https://doi.org/10.1007/s11948-023-00445-1

Steele, L., Lee, H. L., Earp, E., Hong, A., & Thomson, J. (2021). Who writes dermatology randomized controlled trials? The need to specify the role of medical writers [Article]. Clinical and Experimental Dermatology, 46(6), 1086-1088. https://doi.org/10.1111/ced.14711

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. PLOS ONE, 6(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tolle, K. M., Tansley, D. S. W., & Hey, A. J. G. (2011). The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. Proceedings of the IEEE, 99(8), 1334-1337. https://doi.org/10.1109/JPROC.2011.2155130

Treadway, J., Hahnel, M., Leonelli, S., Penny, D., Groenewegen, D., Miyairi, N., Hayashi, K., O'Donnell, D., Science, D., & Hook, D. (2016). The State of Open Data Report. Figshare. https://figshare.com/articles/report/The_State_of_Open_Data_Report/4036398?file=6558051

Udey, M. C. (2018). Giving Credit where Credit Is Due (and Assigning Individual Responsibilities) [Editorial Material]. Journal of Investigative Dermatology, 138(7), 1451-1452. https://doi.org/10.1016/j.jid.2018.05.010

Vasilevsky, N. A., Hosseini, M., Teplitzky, S., Ilik, V., Mohammadi, E., Schneider, J., Kern, B., Colomb, J., Edmunds, S. C., Gutzman, K., Himmelstein, D. S., White, M., Smith, B., O'Keefe, L., Haendel, M., & Holmes, K. L. (2021). Is authorship sufficient for today's collaborative research? A call for contributor roles [Article]. Accountability in Research-Ethics Integrity and Policy, 28(1), 23-43. https://doi.org/10.1080/08989621.2020.1779591

Wallis, J. C., & Borgman, C. L. (2011). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. Proceedings of the American Society for Information Science and Technology, 48(1), 1-10. https://doi.org/https://doi.org/10.1002/meet.2011.14504801188

Wilson, A., Downs, R. R., Lenhardt, W. C., Meyer, C., Michener, W., Ramapriyan, H., & Robinson, E. (2014). Realizing the Value of a National Asset: Scientific Data. Eos, Transactions American Geophysical Union, 95(50), 477-478. https://doi.org/https://doi.org/10.1002/2014EO500006

Wittenburg, P. (2021). Open Science and Data Science. Data Intelligence, 3(1), 95-105. https://doi.org/10.1162/dint_a_00082

Zhang, Z., Wang, S. D., Li, G. S., Kong, G., Gu, H., & Alfon, F. (2019). The contributor roles for randomized controlled trials and the proposal for a novel CRediT-RCT [Article]. Annals of Translational Medicine, 7(24), 812, Article 812. https://doi.org/10.21037/atm.2019.12.96

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

A benchmark for evaluating crisis information generation capabilities in LLMs

Ruilian Han, Lu An, Wei Zhou, and Gang Li
DOI: https://doi.org/10.47989/ir30iConf47518

Abstract

Introduction. Large language models (LLMs) have become increasingly significant in crisis information management due to their advanced natural language processing capabilities. This study aims to develop a comprehensive evaluation benchmark to assess the effectiveness of LLMs in generating crisis information.

Method. CIEeval, an evaluation dataset, was constructed through steps such as information extraction and prompt generation. CIEeval covers 26 types of crises across sub-domains including water disasters, environmental pollution, and others, comprising a total of 4.8k data entries.

Analysis. Eight LLMs applicable to the Chinese context were selected for evaluation based on multidimensional criteria. A combination of manual and machine scoring methods was utilized. This approach ensured a comprehensive understanding of each model's performance.

Results. The manual and machine scores showed significant correlation. Under this scoring method, Claude 3.5 Sonnet performed the best, particularly excelling in complex scenarios like natural and accident disasters. In contrast, while scoring slightly lower overall, Chinese models like ERNIE 4.0 Turbo and iFlytek Spark V4.0, showed strong performance in specific crises.

Conclusion. The evaluation benchmark validates the best LLM for crisis information generation (Claude 3.5 Sonnet) and provides valuable insights for LLMs to optimize and apply LLM in crisis information.

Introduction

In the context of rapid technological development, large language models (LLMs) have become a key focus in natural language processing due to their application of deep learning and Transformer architecture. LLMs efficiently encode and decode language, demonstrating significant advancements in machine translation, text generation, and dialogue systems (Guler et al., 2024).

Crisis information (CI) is increasingly important in crisis response. Traditional manual process is inefficient and hard to manage large-scale and frequent events. LLMs, with their powerful language generation capabilities, can automatically produce high-quality crisis information, improving efficiency and accuracy. Additionally, LLMs handle multilingual and multimodal information, facilitating cross-regional and cross-domain information sharing. However, the information generated by different LLMs varies greatly in quality, and may exhibit over-interpretation of information or thematic bias, both of which may affect the effectiveness of decision-making.

This study aims to establish benchmarks to comprehensively evaluate LLMs' performance in CI generation across scenarios like natural disasters, accidents, and public health incidents. By assessing objectivity, completeness, and reasonableness, this research provides insights into LLMs' strengths and limitations, offering guidance for model optimization and practical applications in CI.

Literature review

In recent years, several specialized benchmark datasets have been developed to assess the performance of LLMs, such as GLUE (Wang et al., 2019), SuperGLUE (Sarlin et al., 2020), and SoEval (Liu et al., 2024). These benchmarks cover tasks like reading comprehension, sentiment analysis, and structured output, becoming essential tools for evaluating LLMs. More complex benchmarks like BIG-Bench Hard have been introduced recently, facilitating the evaluation of LLMs' cross-task and cross-domain generalization capabilities (Suzgun et al., 2022). Evaluation benchmarks in the Chinese context are also emerging, such as CLUE (Xu et al., 2020), which has become a widely used evaluation tool in various industries in China.

In the information science field, LLMs are widely applied in automating information analysis (Giannakopoulos et al., 2023), processing, and public opinion monitoring. As LLMs are increasingly used, researchers have established evaluation frameworks for information compilation and report generation. For example, Thelwall (2024) assessed LLMs' effectiveness in scientific information evaluation, and explored ways to enhance their capabilities through prompt engineering and external tools.

For evaluation methods, while accuracy remains an important measure of LLMs' objectivity, automated metrics like BLEU (Evtikhiev et al., 2023) and Bipol (Alkhaled et al., 2023) have limitations in addressing the feasibility of generated content for open-ended questions. Therefore, manual evaluation remains essential, especially in text generation and translation tasks (Xu et al., 2023).

Currently, there is a lack of specialized LLM benchmarks for crisis information generation. Crisis information, as a core element for responding to and solving crises, requires the support of the latest technologies. Therefore, this paper proposes constructing CIEval, an LLM evaluation benchmark for CI generation, combining manual and machine evaluations to identify the best LLMs and promote their application in this field.

Generation of CIEval dataset

Subjects selection

The crisis mechanism aims to efficiently respond to sudden crises and ensure social stability and public safety. According to the definition of *the Emergency Response Law* (General Office of the State Council, 2024), crises include natural disasters, accident disasters, public health events, and social security events. In order to comprehensively and accurately evaluate the ability of LLMs in CI generation, the CIEval dataset we constructed covers 16 segmented subjects under the four major categories of crises mentioned above, ensuring its representativeness and wide applicability. In addition, to further enhance the completeness of the dataset, we also focus on a specific social security event - cyber security, and includes 10 events with high frequency of occurrence in this field as supplementary categories. The specific event classification is shown in Table 1.

Categories	Subjects	Categories	Subjects	
Natural disasters	Drought and water disasters	Social security events	Economic security incidents	
	Meteorological disasters		Foreign emergencies	
	Earthquake		Cyber security events	Botnet
	Geologic disasters			Data leakage
	Forest and grassland fires			Phishing emails
Accident disasters	Enterprise safety accidents			Vulnerability exploitation
	Transportation accidents			DDOS
	Public facility accidents			APT
	Environmental pollution			Tampering
	Ecological damage incident			Worms
Public health events	Infectious disease outbreaks			Mining
	Congregative unknown disease			Ransomware
	Food safety			
	Animal outbreaks			

Table 1. Subjects of CIEval

Data collection and pre-processing

The primary sources of the raw data are post-disaster recovery plans and accident investigation reports publicly released by emergency management departments. Data related to cyber security events comes from typical case collections issued by cyber security companies like QAX (https://en.qianxin.com/).

For the raw data, a refined content extraction strategy was implemented to remove non-event-related elements such as headers and hyperlinks, retaining and integrating only text and image content crucial for event analysis, which are stored in structured documents. In cases where a document may cover multiple events, detailed manual intervention was employed to carefully split the content. This process ensures that each independent event is accurately mapped to a unique corresponding document, facilitating subsequent dataset construction. Each final document includes an overview, losses, policy measures, and other information of the event.

Dataset generation method

Dataset generation

Advanced models like GPT-4.0 play a crucial role in dataset construction (Liu et al., 2024). Based on this, we utilized the generative capabilities of LLMs to build an evaluation dataset, significantly shortening the construction period and reducing manual workload, providing abundant test resources for assessing LLM performance in specific crisis information needs.

Prompt engineering is a core element focused on designing precise prompts to guide LLMs in producing expected outputs. The key to prompt engineering lies in how to accurately construct prompts. Prompts (*P*) typically consist of instructions (*I*) and inputs (*In*); the instructions specify the task goals, setting a framework for model responses, while the inputs provide specific context or examples. *I* and *In* jointly influence the quality of the model's output, represented by Eq. (1).

$$P = f(I, In) \tag{1}$$

Here, *I* and *In* are combined through function *f* to form the LLM prompt (*P*). In this paper, *f* is the internal mechanism of LLMs, implemented by GPT-4.0. The method of using LLMs to generate datasets is illustrated in Figure 1.

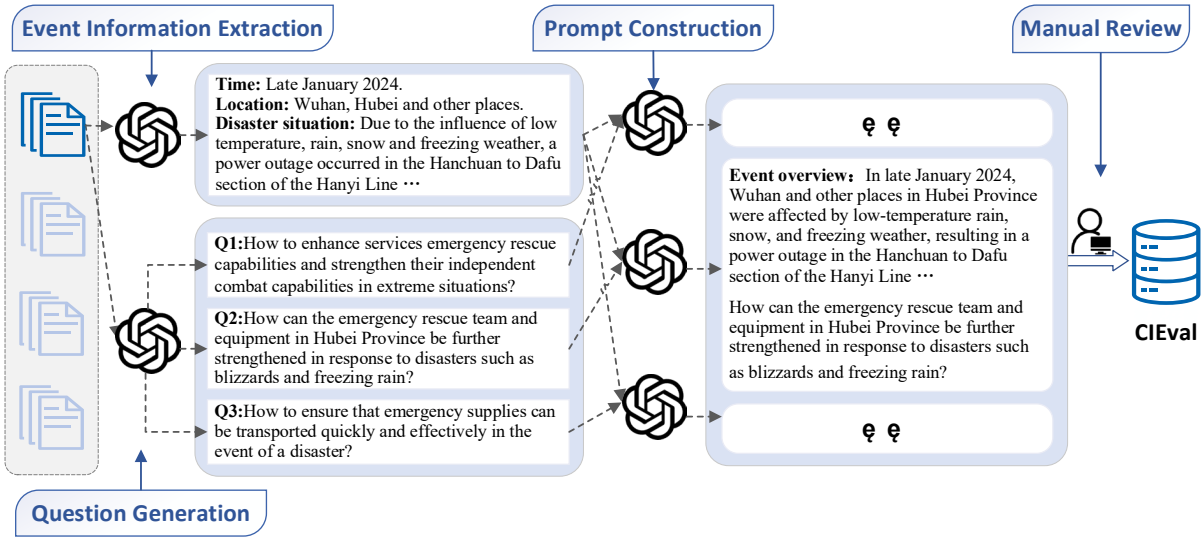


Figure 1. Overview of CIEval generation methods

Specifically, the dataset construction process we adopted follows a systematic pipeline approach, detailed as follows:

1. *Event information extraction*: pre-processed documents are input into GPT-4.0, which, using its powerful analytical capabilities, automatically extracts key event information from the documents. The information includes but not limited to occurrence time, disaster type, and economic losses, etc., i.e., *In*.

2. *Targeted question generation*: based on the response measures, policy protection suggestions, and other information in the input document, GPT-4.0 randomly generates a series of crisis-related questions, i.e., *I*.

3. *Prompt construction*: in this step, GPT-4.0 is used to combine *In* and *I* to form a complete *P* dataset for assessing the CI generation capability of LLMs.

4. *Manual review*: the automatically generated prompts undergo rigorous manual review to ensure clarity, fluency, and high relevance to the crisis information tasks. This process further minimizes the potential influence of any biases introduced by GPT-4.0, thereby enhancing the dataset quality and ensuring the accuracy and validity of the evaluation results.

Avoiding data contamination

In constructing the dataset, this study focuses on avoiding data contamination to ensure quality. Given that large-scale crises, especially those that generate widespread social response online, often become publicly accessible data resources through subsequent investigation reports, there is a potential risk of these being included in large language model training datasets. Considering

that the latest training data for the model to be evaluated is up to April 2024, this paper uses data from April 2024 onwards as the original dataset for CIEval, in order to reduce data contamination.

Dataset description

CIEval is a comprehensive dataset designed to thoroughly evaluate the performance of LLMs in CI generation. The dataset includes four major categories of authoritative-defined crisis types and additionally incorporates cyber security events, which have a broad impact in modern society, to fully reflect realistic and diverse CI needs. It covers a wide range of event types, including but not limited to meteorological disasters, earthquakes, transportation accidents, food safety, and data leakage, providing a thorough test of the models' performance in handling complex and variable CI tasks. Table 2 shows some of the contents of this dataset. CIEval contains a total of 4,820 evaluation questions, offering a comprehensive benchmark framework to assess the overall capabilities of LLMs in CI generation tasks, ultimately enhancing their practicality and effectiveness in real-world crisis scenarios.

Categories	Questions	Examples
Natural disasters	1500	**Event overview:** 'Time: June 22 to July 4, 2024; Location: Jiangxi Province; Affected Areas: 65 counties (cities, districts) in 9 prefecture level cities including Nanchang and Jiujiang; Number of Affected Persons: 1.565 million people...' **Based on the above description, please answer the following:** How to coordinate multi departmental cooperation to ensure the basic living security of residents in disaster areas?
Accident disasters	2000	**Event overview**: 'Time: April 26, 2024; Location: Xia County Expressway in Yuncheng, Shanxi Province; Event Type: Vehicle Fire Caused by Traffic Accident...' **Based on the above description, please answer the following:** How to handle the automatic locking system of damaged vehicles during crisis rescue?
Public health events	550	**Event overview:** '5.27 Xinzheng Elementary School Canteen Food Mold Incident' **Based on the above description, please answer the following:** How to effectively establish and manage a parental supervision mechanism in the school environment to ensure food safety?
Social security events	770	**Event overview:** '8.19 Philippine Coast Guard Ship Collides with Chinese Ship Incident' **Based on the above description, please answer the following:** How to quickly take measures to prevent the situation from escalating when a similar collision event occurs?
Cyber security events	310	**Event overview:** 'The terminal computer was infected with ransomware through phishing emails' **Based on the above description, please answer the following:** What role does terminal security control software play in preventing ransomware?

Table 2. Partial content of CIEval dataset

Benchmarking experiment

Models

To comprehensively understand the applicability and CI generation capabilities of LLMs in the CI domain within the Chinese context, this study selected four of the latest models developed by Chinese companies and four internationally renowned models suitable for the Chinese context. These models vary in scale and structure, performing differently across various datasets and tasks. Specific information about the models is shown in Table 3.

Models	Developer	Size	Access	Source
GPT-4o	OpenAI	Undisclosed	API	https://chat.openai.com/
Claude 3.5 Sonnet	Authropic	Undisclosed	API	https://claude.ai
Llama-3.1-405B	Meta	405B	Weights	https://github.com/meta-llama/llama3
Gemini-1.5-Pro	Google	Undisclosed	API	https://deepmind.google/technologies/gemini/
ERNIE 4.0 Turbo	Baidu	Undisclosed	Official website	https://yiyan.baidu.com
GLM-130B	Tsinghua	130B	Weights	https://github.com/THUDM/GLM-130B
iFlytek Spark V4.0	iFlytek	Undisclosed	Official website	https://xinghuo.xfyun.cn
Qwen LM-72B	Alibaba	72B	Official website	https://tongyi.aliyun.com/qianwen

Table 3. Description of models

GPT-4o is the latest and fastest flagship model of OpenAI, a specialized version of GPT-4.0 optimized for specific tasks such as real-time inference, and generating audio, images, and text. In addition, Claude 3.5 Sonnet surpassed GPT-4o in multiple areas including graduate level reasoning, undergraduate level knowledge, and coding ability, according to the Authropic's release report.

ERNIE 4.0 Turbo, GLM-130B, IFlytek Spark V4.0 and Qwen-72B are currently outstanding Chinese LLMs. IFlytek Spark V4.0 fully benchmarked GPT-4-Turbo during construction, and according to testing, it has surpassed GPT-4-Turbo in text generation, logical reasoning, and other aspects. Among them, non-open-source models are all the premium version.

Evaluation methods

This study plans to use a combined evaluation method of manual scoring and machine scoring to enhance the efficiency of the evaluation process and ensure the comprehensiveness and accuracy of the results.

Manual scoring

Based on Wang et al. (2024), we established evaluation indicators for CI generation capabilities of LLMs, including content quality, expression quality, feasibility, and effectiveness, starting from the quality of the CI itself. Firstly, in order to obtain professional evaluation opinions, this study invited three PhD students with relevant experience in the field of CI as evaluators. The scoring personnel score the CI generated by each model based on the indicators, and the score level is divided into [extremely high, high, medium, low, extremely low]. Secondly, based on the fuzzy decision trial and evaluation laboratory method, the weights of each indicator were based on the experience of the scoring personnel. Finally, the interactive multi-criteria decision-making method based on triangular intuitive fuzzy numbers was used to obtain the global dominance of the evaluated model, which is the CI generation capability score of LLMs (Qin et al., 2017).

Machine scoring

Given the large scale of CIEval, relying solely on manual scoring would require a significant amount of manpower and resources. To this end, this study further explores the possibility of machine scoring. We select GPT-4.0, which performs well in multiple benchmarks (Xu et al., 2023), and Claude 3.5 Sonnet, which the developer claims to be superior to GPT-4o in all aspects, as '*machine scoring experts*.'

Similar to the final results obtained through manual scoring, the machine scoring range is set to [0,1]. To ensure the rationality of machine scoring, we will conduct validation experiments before

formal evaluation, comparing the consistency between GPT-4.0 and Claude-3.5 Sonnet scoring and manual scoring to verify their reliability as machine scoring experts. Choose the optimal model score to replace manual scoring, thereby improving evaluation efficiency and reducing labour costs.

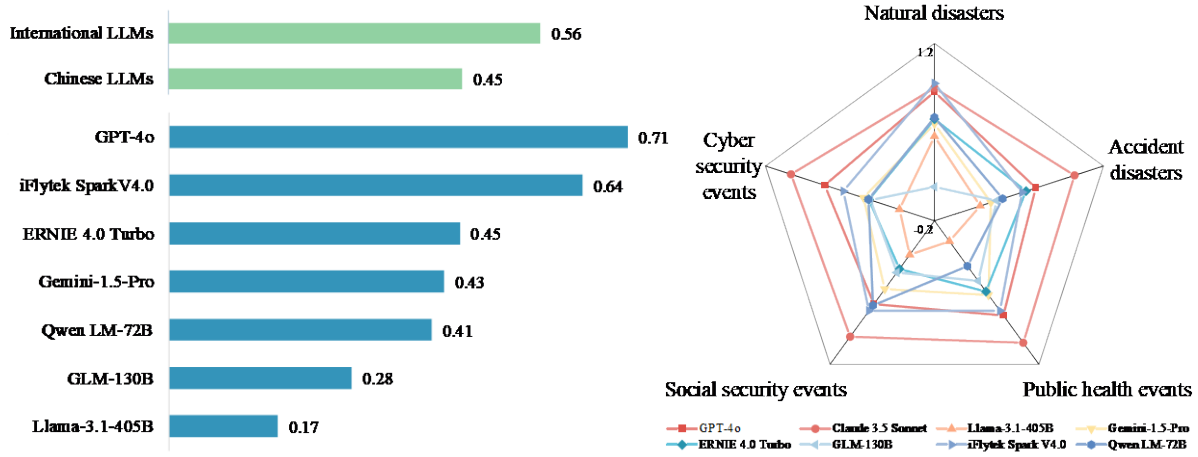
Results and discussion

Comparison between manual scoring and machine scoring

This study randomly selected 80 pieces of data in the CIEval dataset as samples for manual scoring and machine scoring, and conducted correlation tests between manual scoring and Claude-3.5 Sonnet/GPT-4.0 scoring, respectively. The Mann-Kendall's test results showed that the correlation coefficient between GPT-4.0 score and the manual score was 0.397, showing extremely high significance ($p<0.01$). Claude-3.5 Sonnet was significantly correlated with manual scoring at the p value lower than 0.05. To further validate the rationality of utilizing GPT-4.0 for scoring, a Spearman test was conducted. The results indicated a correlation coefficient of 0.514 between GPT-4.0 scores and manual scores, further supporting the feasibility and reliability of replacing manual scoring with GPT-4.0. Therefore, this study selects GPT-4.0 as the main tool for evaluating CI generation capabilities of LLMs.

The crisis information generation capabilities of LLMs

This section provides a performance evaluation of several LLMs based on the CIEval dataset. Overall, Chinese LLMs slightly underperform compared to leading international models such as the GPT and Claude series (Figure 2(a)). This performance gap can be attributed to the global advantages in data resources and technological development enjoyed by international models.



(a) Average score of LLMs' CI generation capability

(b) Score of LLMs in different types of crisis

Figure 2. Score of LLMs' CI generation capability

As shown in Figure 2(b), iFlytek Spark V4.0 excels in generating relevant information for natural disasters, achieving high scores due to its deep understanding of local meteorological data. In the context of accident disasters, Claude 3.5 Sonnet and GPT-4o outperform others with scores of 0.96 and 0.63, respectively. While models like ERNIE 4.0 Turbo perform well in specific scenarios such as transportation accidents (Figure 3), others, like Llama-3.1-405B, fall short in complex enterprise safety incidents, producing less actionable information. For public health events, Claude 3.5 Sonnet leads in scenarios like infectious disease outbreaks. In the social security category, Claude 3.5 Sonnet excels, achieving a score of 0.98 due to its strong situational analysis. Other models like Qwen LM-72B and ERNIE 4.0 Turbo achieve high scores in specific incidents like mining and ransomware.

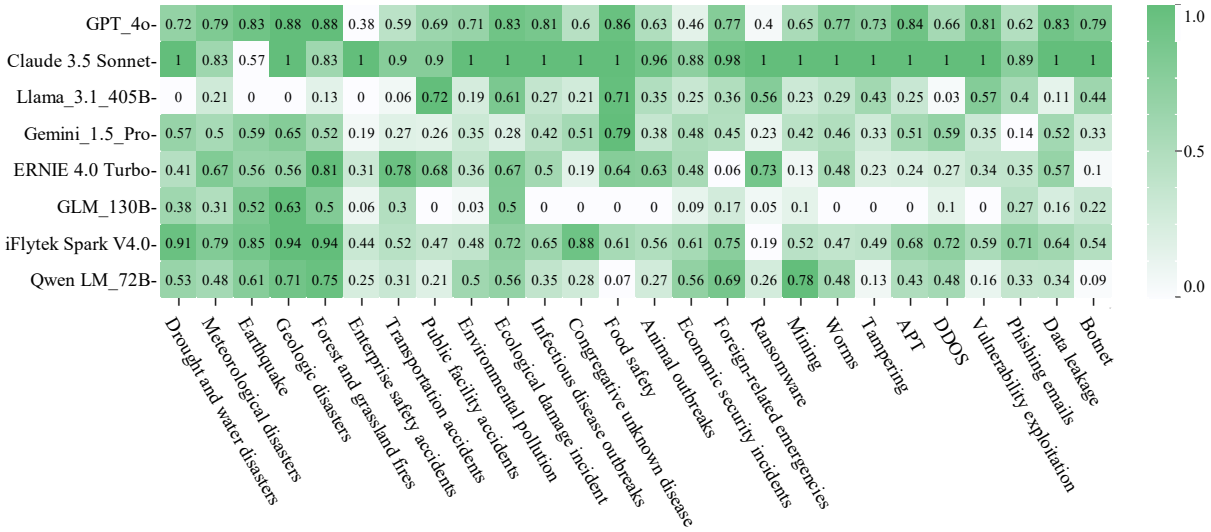


Figure 3. Score of LLMs' CI generation capability in segmented scenarios

Conclusion

This study focuses on the potential of LLMs in the field of CI generation and proposes a scientifically reasonable evaluation benchmark, CIEval. CIEval is constructed through a series of processes including key event information extraction, question generation, prompt construction, and manual review. This dataset contains 26 types of crises related to natural disasters, accident disasters, public health events, and social security events, with a total of 4.8k data. In the experimental phase, we evaluated the CI generation capability of LLMs such as Claude 3.5 Sonnet using the CIEval dataset. At the same time, we validated the feasibility of GPT-4.0 as a machine scoring expert. This benchmark aims to provide reference for the optimization and application of LLMs in CI generation in the future, which will help promote the application and development of LLMs in practical emergency management.

Acknowledgements

This study is supported by the National Social Science Foundation of China (Grant No. 23&ZD230).

About the authors

Ruilian Han is a PhD student at School of Information Management, Wuhan University, China. Her research focuses on social media data analysis. She can be contacted at rlhan_1127@163.com

Lu An is a professor at School of Information Management, Wuhan University, China. Her research focuses on crisis informatics. She can be contacted at anlu97@163.com

Wei Zhou is a PhD student at School of Information Management, Wuhan University, China. Her research focuses on risk identification. She can be contacted at 664880781@qq.com

Li Gang is a professor at School of Information Management, Wuhan University, China. His research focuses on information resource management. He can be contacted at ligang@whu.edu.cn

References

Alkhaled, L., Adewumi, T., & Sabry, S. S. (2023). Bipol: A novel multi-axes bias evaluation metric with explainability for NLP. Natural Language Processing Journal, 4, 100030. https://doi.org/10.1016/j.nlp.2023.100030

Evtikhiev, M., Bogomolov, E., Sokolov, Y., & Bryksin, T. (2023). Out of the bleu: How should we assess quality of the code generation models? Journal of Systems and Software, 203, 111741. https://doi.org/10.1016/j.jss.2023.111741

General Office of the State Council, P. (2024). Emergency Response Law. People's Publishing House.

Giannakopoulos, K., Kavadella, A., Salim, A. A., Stamatopoulos, V., & Kaklamanos, E. G. (2023). Evaluation of the performance of generative AI large language models Chatgpt, Google bard, and microsoft bing chat in supporting evidence-based dentistry: comparative mixed methods study. Journal of Medical Internet Research, 25, e51580. https://doi.org/10.2196/51580

Guler, N., & Kirshner, S.N. (2024). A literature review of artificial intelligence research in business and management using machine learning and ChatGPT. Data and Information Management, 8(3), 1-25. https://doi.org/10.1016/j.dim.2024.100076

Liu, Y., Li, D., Wang, K., Xiong, Z., Shi, F., Wang, J., Li, B., & Hang, B. (2024). Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. Information Processing & Management, 61(5), 103809. https://doi.org/10.1016/j.ipm.2024.103809

Qin, Q., Liang, F., Li, L., Chen, Y.-W., & Yu, G.-F. (2017). A TODIM-based multi-criteria group decision making with triangular intuitionistic fuzzy numbers. Applied Soft Computing, 55, 93–107. https://doi.org/10.1016/j.asoc.2017.01.041

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4938–4947. https://doi.org/10.1109/CVPR42600.2020.00499

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2022). Challenging Big-Bench tasks and whether chain-of-thought can solve them (arXiv:2210.09261). arXiv. https://doi.org/10.48550/arXiv.2210.09261

Thelwall, M. (2024). Can ChatGPT evaluate research quality? Journal of Data and Information Science, 9(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding (arXiv:1804.07461). arXiv. https://doi.org/10.48550/arXiv.1804.07461

Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S., & Aggarwal, S. (2024). Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality. Journal of the Knowledge Economy, 15(1), 1159–1178. https://doi.org/10.1007/s13132-022-01096-6

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., … Lan, Z. (2020). CLUE: A chinese language understanding evaluation benchmark (arXiv:2004.05986). arXiv. https://doi.org/10.48550/arXiv.2004.05986

Xu, L., Li, A., Zhu, L., Xue, H., Zhu, C., Zhao, K., He, H., Zhang, X., Kang, Q., & Lan, Z. (2023). SuperCLUE: A comprehensive chinese large language model benchmark (arXiv:2307.15020). arXiv. https://doi.org/10.48550/arXiv.2307.15020

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Design of audio ads to prebunk misinformation and promote civil discourse

Diana Daly and Kainan Jarrette
DOI: https://doi.org/10.47989/ir30iConf47359

Abstract

Introduction. This paper reports on an experimental pilot study, involving 99 student participants at a large southwestern US university, that measured the effectiveness of ads designed to prebunk, or increase scepticism, of audio misinformation to increase information literacy.

Method. The research team created short prebunking advertisements based on audio recordings of theatrical sketches they developed, some further enhanced by AI-generated visualizations of the transcripts. The team then used new and existing measurement tools to gauge the effectiveness of one of the ads in increasing information literacy.

Results. Prebunking ads were found to be effective in increasing information literacy overall, with ads involving AI-assisted visualization correlated with improved information literacy, and audio-only ad findings more mixed.

Introduction

In the face of escalating misinformation and radicalization in digital audio, our research team sought to develop and evaluate a novel intervention method designed to prebunk problematic audio content before it takes hold individually or socially and enhance our approach using artificial intelligence for speed and scalability. Our study focused on the creation and deployment of short prebunking audio advertisements, some further enhanced by AI-generated visualizations of the transcripts, to assess their effectiveness in building information literacy gauged through identifying true and false news headlines.

Background

Audio-based misinformation and radicalization are topics of increasing concern in digital environments (Muhammed T & Mathew, 2022; Wirtschafter 2023), where lies spread faster than truth (Vosoughi et al., 2018) and can cause serious harm. While there is a canon of new research on media misinformation interventions, audio misinformation and polarization interventions have not been sufficiently studied. Pew Research has found many Americans listen to and get information from audio sources, including podcasts which have gained popularity over the previous decade, and terrestrial radio which has continued to reach large audiences (Jurkowitz, 2023). Research on podcasts in particular has found high trust in hosts and advertisements (Brinson & Lemon, 2023), and significant spread of misinformation. The affordances of audio-centric content are well-suited for manipulation: voice-based misinformation and radicalization in Lebanon were found to exploit the medium's conveying of emotion and speaker source credibility (El Masri et al 2022). Of the scant research on audio misinformation interventions, one salient study successfully deployed audio dramas designed by researchers to debunk common health misconceptions among listeners in Sierra Leone (Winters et al 2021).

One strategy of increasing interest among misinformation interventions is prebunking, or the act of trying to build mental defences against manipulation techniques before problematic content is seen or consumed. Prebunking typically involves presenting consumers with a pre-emptive warning about the tactics used in misinformation, coupled with a refutation before they encounter the false information itself (Roozenbeek & van der Linden 2019). In some studies, prebunking has been found to be significantly more effective than debunking (Banas & Miller 2013; Lewandowsky & van der Linden, 2021). A foundation of prebunking is McGuire's (1961) Inoculation Theory, positing that exposing individuals to a weakened form of an argument along with refutations can build cognitive resistance against misinformation (McGuire 1964).

Our current work focuses on prebunking because this approach has shown promise in the current media landscape, where rapid misinformation dissemination often outpaces the ability to correct falsehoods effectively. To create prebunking messages, members of our research team drew on our backgrounds in audio production and theatre including sketch and improvisation (improv) to script, perform, record, and edit short audio sketches designed to be played as ads during podcasts. In acknowledgment of the growth of video podcasting and of limitations in our capacity to create video ads, we also used AI to add visualization to the ads we created. Here we report on our ad creation processes and our experimental pilot study conducted to gauge one ad's influence on information literacy.

Methods

Our research team created five audio ads that humorously dramatized misinformation- or radicalization-related situations and modelled civil responses. We next visualized the audio transcripts using artificial intelligence. The program descript provided general visual templates that animated words being spoken, with distinct colour schemes and visual waveforms for each speaker in the ad. Our team then finessed visualizations through minor changes in item placement and colour in video editing software. We also conducted experimental research to learn about the

effectiveness and impacts of one of the ads as an experimental pilot study. In this section, we explain the ad creation and study creation processes.

Ad title	Description	Targeted misinformation	Ad strategies
Buy This Idea!	Two men consider buying products in the Marketplace of Ideas, and one narrowly escapes a bad outcome.	marketplace of ideas; influencer-spread misinformation; truth from sport; polarization	masculine identification; vocal techniques; scatological humour; compassion
New Friends	Mel reunites with an old friend, but when she describes her new friends, he offers an alternative.	participatory disinformation; misinformation; radicalization	neutral location; vocal techniques; musical techniques; reward of community
Zoo Moms	Two moms and their kids meet at the zoo, where a new health trend is tested in the bear den.	health misinformation; wellness trends; distrust in medical science	feminine identification; motherhood identification; voice play; compassion
Dog Show	Two friends meet at the Dog Show, and a vendetta against schnauzers is exposed.	the ends justify the means; media manipulation; polarization	nonthreatening target proxy; emotional identification; compassion
Heroes and Sidekicks	Gas Man the superhero is confronted with the unintended effects of his powers, until his sidekick offers a compromise.	glorification of the past; motivated reasoning; science manipulation	masculine identification; flatulence humour; friendship; vocal techniques; sound design techniques

Table 1. Ads designed

Ad Design

The five ads we created all humorously dramatized situations in which people interact when misinformation or manipulation is involved, resolving when the problem is managed civilly. All ads end with a reference to our project website to learn more. Our first ad took three months to create, after which we began creating ads in one month or less in overlapping phases. Our ad creation process included steps described in this section.

Identifying misinformation to prebunk

Our first step in creating each prebunking ad was to identify a misinformation strategy or manipulation technique causing continuous harm. Working with a team of students, we analysed

social media trends and common misinformation tactics. One familiar manipulation technique emerging from our analysis was the *marketplace of ideas* (Kim & Gil De Zúñiga 2021), which we made into an actual marketplace in one ad entitled "Buy this Idea!" Our ad "New Friends" used in this study emerged from discussions about QAnon and its influence on the January 6th insurrection to target radicalization, illustrating how misinformation and the reward of community membership can lead to extreme ideological shifts, and how social connections can respond.

Scenario development

Once a misinformation strategy was identified, we brainstormed scenarios that could humorously and effectively engage with the strategy and model non-reactionary responses. Our team used techniques used from improv to manifest a scene through two characters and rapidly establish their relationship and the setting, so the remainder of the ad could be devoted to the attitudes actions and that would model misinformation-related dynamics.

Character development

We adhered to a basic rule of featuring only two main characters with distinct personalities in most ads, with any additional characters serving peripheral roles. This approach was designed around the original audio-only format of the ads to prevent listener confusion. Characters knew each other to model how individuals can engage with friends or family members who may be susceptible to misinformation.

Setting

Some scenarios were grounded in real-world locations or situations, such as the '*New Friends*' ad set in a post office or the '*Zoo Moms*' ad set in a zoo, as familiar and ideologically neutral locations to avoid triggering defensive reactions from the audience. For others we created metaphorical settings such as the marketplace of ideas.

Media production

We designed media to be 60 to 90 seconds in length, to optimize audience attention and allow for eventual podcast ad airing at reasonable cost. Production began with script drafting, followed by recording and editing.

Script drafting

With the scenario and characters established, one team member would take the lead on drafting the script. Subsequent revisions were then collaboratively discussed and refined, with a focus on maintaining the ad's clarity and humour while aligning with the intended time constraints.

Recording

The recording process was designed to maximize flexibility and adaptability. We often booked more voice actors than there were characters, allowing us to create multiple versions of each ad. This approach enabled us to test different voice actor combinations and enables testing on which versions resonate most with listeners. During the recording sessions, we aimed to capture at least one take that adhered closely to the script, while also allowing for improvisation and variations—a technique known as '*retroscripting*' in film and television production.

We provide an example of a transcript, with timestamps and emphases added after recording, for the ad '*New Friends*' used in the pilot study. Table 1 describes all ads created.

[00:00:00] Jason: Oh, hey Mel. Funny running into you here at the post office. We haven't seen you in a while.

[00:00:05] Mel: Well, I've been really busy with my *new* friends. They're pretty great and they *really* get me. Not like this POST OFFICE. They have PROBLEMS!

[00:00:13] Jason: Whoa, whoa, whoa, whoa. That's, that's cool you have new friends. So, do you guys talk karaoke? I know that's kind of your thing.

[00:00:19] Mel: No, but we have some GREAT online chats about Big Postal. Have you seen the price of stamps lately?

[00:00:25] Jason: Uh, well, uh, do you all go bowling? I know how much you love the lanes.

[00:00:29] Mel: They're not really big bowlers. They're more serious than that. Did you know that CARRIER PIGEONS are actually FASTER than MAIL TRUCKS???

[00:00:36] Jason: I, I don't think that's true, but um, hey, have they heard you do The your *amazing* cover of Forest for the Trees?

[00:00:42] Mel: No, but we have talked about storming this post office tomorrow night.

[00:00:48] Jason: Oh boy. Uh, well look, I'm really sorry I haven't called in a while. We all love having you around, so why don't you join us for some karaoke tomorrow night instead? We can talk about whatever you want.

[00:00:59] Mel: Really? [00:01:00] Well, I guess we could just *not* storm the post office. I mean, it *is* conveniently located.

[00:01:13] Narrator: It's great to find a new community online. But you shouldn't have to be extreme just to get them to accept you. Find out more at immersivetruth.org.

Audio editing

Different versions of the ads were created through editing using various combinations of voice actors, with a priority placed on meeting the target duration without compromising the message's clarity.

Intended uses and applications of Ads

Ads created in this project are designed to be aired during podcasts, particularly those known to spread misinformation. Our work focused on audio prebunking is modelled in part on other campaigns designed to intervene in misinformation spread. Previously researched prebunking video campaigns have been scaled up and distributed broadly to mitigate the harms of misinformation in locations including Europe (Lewandowski and Van der linden, 2021) and India (Parihar et al, 2024). While our campaigns on audio misinformation interventions could be similarly scaled up with corporate partnership, we intend at the time of this writing to scale up more modestly, through modeling collaborative and user-based ad creation in workshops and educational settings, and through production accelerated using AI assistance.

Experimental study

The primary goal of our study was to evaluate the effectiveness of our prebunking ads in reducing susceptibility to misinformation, increasing the ability to identify false news headlines, and influencing attitudes towards civil discourse. We formulated specific research questions to guide our inquiry thus far:

1. Will exposure to a specific prebunking ad enhance an individual's ability to identify false news headlines?
2. How will the visualization of ad script using AI assistance impact respondent literacy?

3. How will participants describe and gauge the effectiveness of a specific prebunking audio ad?

These questions framed our study design, which involved a controlled experimental setup with undergraduate students at a large university in the southwestern US as participants.

Study participants and recruitment

We recruited participants through instructors of courses related to study topics, encouraging the offer of extra credit as an incentive to participate. Recruitment materials included flyers and wording for announcements to be included in the course management system D2L Brightspace used by the university.

Data collection

Participants who followed the invitation link provided through the recruitment materials first completed a pre-screening survey to assess their eligibility. Eligible participants were then invited to participate in Part 1 of the study, which involved providing informed consent, demographic information, and details about their social media practices.

In addition to answering questions about themselves and their social media use, participants completed qualitative questionnaires and three scales: the generic conspiracist belief scale (Brotherton et al., 2013), the misinformation susceptibility test or MIST-20 (Maertens et al., 2024). We also developed and applied our own misinformation susceptibility test based around the model of the MIST-20, called ITMIST, using different real and fake headlines.

Participants were then exposed to one of the prebunking ads or a control ad, in a format of either audio-only or an audiogram with AI-generated visualizations. Following ad exposure, participants completed a survey response questionnaire, marking the end of Part 1. The next day, participants received a link to Part 2 of the study, where they were asked to take the GCBS, MIST-20, and ITMIST scales and completed another qualitative questionnaire within one week of their completion of the first survey.

Results

Demographic information and social media practices

143 respondents filled out Part 1 of the survey; of these, 99 went on to complete Part 2 of the survey within one week as was required. Participants were predominantly young adults, with ages ranging between 18 and 48. 59 identified as female (59.6%), 38 as male (38.4%), 2 preferred not to disclose their gender (2.1%), and none chose other gender options. 54 identified as White/Caucasian (54.5%), 9 as Hispanic (9.1%), 8 as Asian/Pacific Islander (8.1%), 5 as Black/African American (5.1%), 2 as American Indian/Alaskan Native (2.0%), 19 as multiracial or from other backgrounds (19.2%), and 2 preferred not to disclose their ethnicity (2.0%). Politically, 34 participants identified as Democrats (34.4%), 32 as Republicans (32.3%), 18 as Independents (18.2%), and 15 affiliated with other parties or chose not to disclose political preference (15.6%). In terms of religious affiliation, 37 identified as Christian (37.4%), 16 as non-Religious but Spiritual (16.2%), 11 as Roman Catholic (11.1%), 6 as Jewish, 6 as Atheist (6.1% each), 5 as Muslim (5.1%), 3 as Agnostic, 3 as Protestant (3% each), 5 as another affiliation (5.1%), and 6 preferred not to say (6.1%).

100% of respondents stated that they used a social media platform at least once a week. Of the major platforms, 89 respondents use Instagram (89.9%), 67 use Tik Tok (67.6%), 66 use Snapchat (66.7%), 34 use Twitter/X (34.3%), 21 use Facebook (21.2%), and 8 use other platforms (8.1%). In terms of average daily time spent on social media, 43 respondents reported usage over 2 hours a day (43.4%), 26 between 90 minutes and 2 hours (26.3%), 12 between 60-90 minutes (12.1%), 14 between 30-60 minutes (14.1%), and 4 less than 30 minutes (4.0%). When asked how much of their information on current events/politics/news respondents get from social media, 28 said Most (28.3%), 67 said Some (67.7%), and 4 said None (4.0%).

For podcast engagement, 23 respondents said they frequently tune in to podcasts (23.2%), 50 said sometimes (50.5%), and 26 said never (26.3%). Of respondents who listen to podcasts, 52 reported they always skip podcast ads (71.2%), 19 said they sometimes skip them (26.0%), and 1 said they never skip them (1.4%). There was a wide range of podcasts that respondents said they listened to for entertainment and education, and clustered most strongly by political affiliation. Of respondents who identified as Democrats, the most popular podcasts for entertainment were '*Call Her Daddy*' (CHD) and '*Cancelled*', while the most popular answer for educational podcasts was '*None*'. Of respondents who identified as Republicans, '*The Joe Rogan Experience*' (JRE) was the most popular podcast both for entertainment and for education. Of Independents and those who preferred not to specify their political affiliation, the most popular entertainment podcasts were JRE and CHD, and for education the most common response was '*None*' followed by JRE.

MIST-20 and ITMIST averages

Between the Visual Control group and the Visual Experimental group, the Experimental group showed a significantly larger average improvement on the MIST-20 (Visual Control $\bar{x} = 0.33$, Visual Experimental $\bar{x} = 0.93$), and a moderate improvement on the ITMIST (Visual Control $\bar{x} = 0.81$, Visual Experimental $\bar{x} = 0.98$).

Between the Audio Control group and the Audio Experimental group, there were more mixed results. The Control group showed higher improvement than the Experimental group on the MIST-20 (Audio Control $\bar{x} = 1.41$, Audio Experimental $\bar{x} = 0.85$), but lower improvement than the Experimental group on the ITMIST (Audio Control $\bar{x} = 0.45$, Audio Experimental $\bar{x} = 0.78$).

Individual respondent improvements and declines

Viewed by group, respondents who improved across the two tests showed a larger change as an absolute value than respondents who showed a decline. Subsequently, results leaned toward improved averages across all groups.

Other materials and variables

Results of the GCBS are still being analysed, but do not appear to show clear findings. No significant differences were found based on demographic variables.

Average Improvement Across Groups and Materials

	MIST-20		ITMIST	
	Experimental	Control	Experimental	Control
Visual	**↑0.93**	**↑0.33**	**↑0.81**	**↑0.98**
Audio	**↑0.85**	**↑1.41**	**↑0.45**	**↑0.78**

Percentage Improvement Across Groups and Materials

	MIST-20		ITMIST	
	Experimental	Control	Experimental	Control
Visual	**↑52%**	**↑48%**	**↑62%**	**↑52%**
Audio	**↑56%**	**↑73%**	**↑52%**	**↑45%**

Percentage Decline Across Groups and Materials

	MIST-20		ITMIST	
	Experimental	Control	Experimental	Control
Visual	**↓14%**	**↓43%**	**↓7%**	**↓14%**
Audio	**↓11%**	**↓18%**	**↓22%**	**↓23%**

Table 2. Limited results

Discussion

Our study was designed around audio misinformation interventions and did not emphasize visualization, but results of our early analysis remind us of the axiom popularized by Marshall McLuhan (2017) that '*the medium is the message*.' The ads our team created differed significantly in effectiveness by format: our visualized ads were found to be correlated with improved literacy scores after viewing, while the effectiveness of our audio ads on literacy was not clear. The improvements from the visual ad spanned multiple measures: a greater percentage of respondents in the visual experimental group improved their scores on the MIST-20 and ITMIST compared to the visual control group, and a smaller percentage of respondents in the visual experimental group saw a decline in their scores on those tests when compared to the visual control group.

Results so far show less promise in our audio ad campaign, a finding we attribute in part to increased sensory engagement of video over audio online, and in part to today's online media ecosystem. In light of the increased engagement visuality demonstrates across platforms (Li & Xie, 2020; Dhaoui & Webster, 2021; Sprejer et al., 2022), we theorize the audio-only format left a sensory gap during the playback—and even in deciding whether or not to play the audio—for a population immersed in visual media. The audio-only ads also suffer from a scarcity of genres of popular aural media: while the visualized sketches our team produced may fit into the genre of short videos popular on platforms like TikTok, no elective audio format easily corresponds to our audio-only content, so these are understood as ads. While podcast advertising has been found to heavily

influence listeners, this influence is most noted when the ad is read by the host with whom the listener identifies (Brinson and Lemon 2023), which does not fit the format of our ads.

Future work

Based on early analysis, we see promise in the use of theatrical video sketches to prebunk misinformation, and research planned in the future will include broader populations using all ads created. We plan to increase ad production with AI and to explore their more robust options such as visual storyboarding. While these results show more promise for visualized prebunking ads than audio-only, we judge further work in both types of interventions to be essential, based on our respondents' reliance on podcasts for education and on the spread of audio-only misinformation. Audio campaigns are also more scalable intervention models based for our team, with quality video content taking more time and resources to produce. In future audio-only work, we hope to track playback and move the delivery method offscreen, to capture impacts on listeners performing low-focus tasks like exercise or housework.

Limitations

Limitations include that this pilot study gauged the impact of only one ad described here. Additionally, participants may have learned from the study when they showed improved literacy in the second survey. Most of the population of undergraduate university students completed the study for extra credit in courses unaffiliated with the research team, and some may have entered random answers to earn the credit more quickly. Analyses of results are limited at the time of this writing, with some data including both quantitative and qualitative responses still being analysed. Finally, it would benefit the field of information sciences to develop more tools to measure the effectiveness of these and other misinformation interventions in the future, as measuring the accuracy of headlines may not provide sufficient nuance to accurately measure an individual's scepticism or media literacy.

Acknowledgements

This work was supported by funding through the Foundation and the Vertically Integrated Projects program at the University of Arizona.

About the authors

Diana Daly received her Ph.D. from the iSchool at the University of Arizona. She specializes in Inclusive education in information science; artificial intelligence and society; misinformation and civil discourse; open educational resources and open licensing; digital cultures, and collaboration; Qualitative research methods; new media analysis and production. Dr. Daly is an associate professor in the College of Information Science at the University of Arizona, and currently serves as the college Associate Dean of Undergraduate Academic Affairs and Student Success. She can be reached at didaly@arizona.edu

Kainan Jarrette is a research assistant and current undergraduate student at the University of Arizona. He can be contact at kainanjarrette@arizona.edu

References

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, 39(2), 184–207. https://doi.org/10.1111/hcre.12000

Brinson, N. H., & Lemon, L. L. (2023). Investigating the effects of host trust, credibility, and authenticity in podcast advertising. *Journal of Marketing Communications*, 29(6), 558-576.

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring Belief in Conspiracy Theories: The Generic Conspiracist Beliefs Scale. *Frontiers in Psychology*, 4, 279. https://doi.org/10.3389/fpsyg.2013.00279

Dhaoui, C., & Webster, C. M. (2021). Brand and consumer engagement behaviours on Facebook brand pages: Let's have a (positive) conversation. International Journal of Research in Marketing, 38(1), 155–175. https://doi.org/10.1016/j.ijresmar.2020.06.005

El-Masri, A., Riedl, M. J., & Woolley, S. (2022). Audio misinformation on WhatsApp: A case study from Lebanon. Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-102

Ferreira Caceres, M. M., Sosa, J. P., Lawrence, J. A., Sestacovschi, C., Tidd-Johnson, A., Rasool, M. H. U., Gadamidi, V. K., Ozair, S., Pandav, K., Cuevas-Lou, C., Parrish, M., Rodriguez, I., & Fernandez, J. P. (2022). The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health*, 9(2), 262–277. https://doi.org/10.3934/publichealth.2022018

Gupta, C., Jogdand, D. S., & Kumar, M. (2022). Reviewing the Impact of Social Media on the Mental Health of Adolescents and Young Adults. *Cureus*, 14(10), e30143. https://doi.org/10.7759/cureus.30143

Harjani, T., Roozenbeek, J., Biddlestone, M., van der Linden, S., Stuart, A., Iwahara, M., Piri, B., Xu, R., Goldberg, B., & Graham, M. (2022). A Practical Guide to Prebunking Misinformation. *Interventions*.https://interventions.withgoogle.com/static/pdf/a_practical_guide_to_prebunking_misinformation.pdf

Jerit, J., & Zhao, Y. (2020). Political Misinformation. *Annual Review of Political Science*, 23(1), 77–94. https://doi.org/10.1146/annurev-polisci-050718-032814

Jurkowitz, E. S., Jacob Liedke, Katerina Eva Matsa, Michael Lipka and Mark. (2023, April 18). Podcasts as a Source of News and Information. Pew Research Center. https://www.pewresearch.org/journalism/2023/04/18/podcasts-as-a-source-of-news-and-information/

Kim, J.-N., & Gil De Zúñiga, H. (2021). Pseudo-Information, Media, Publics, and the Failing Marketplace of Ideas: Theory. *American Behavioral Scientist*, 65(2), 163–179. https://doi.org/10.1177/0002764220950606

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 13(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2013). The role of conspiracist ideation and worldviews in predicting rejection of science. *PloS One*, 8(10), e75637. https://doi.org/10.1371/journal.pone.0075637

Lewandowsky, S., & van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. European Review of Social Psychology, 32(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. Journal of Marketing Research, 57(1), 1–19. SAGE Publications Inc. https://doi.org/10.1177/0022243719881113

Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., He, J., & van der Linden, S. (2024). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, 56(3), 1863–1899. https://doi.org/10.3758/s13428-023-02124-2

McGuire, W. J. (1964). Inducing Resistance to Persuasion. Some Contemporary Approaches. *C. C. Haaland and W. O. Kaelber (Eds.), Self and Society. An Anthology of Readings, Lexington, Mass. (Ginn Custom Publishing) 1981, Pp.* 192-230. https://opus4.kobv.de/opus4-Fromm/frontdoor/index/index/docId/16094

McLuhan, M. (2017). The medium is the message. In *Communication theory* (pp. 390-402). Routledge.

Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: A review of research in social media. *International Journal of Data Science and Analytics*, 13(4), 271–285. https://doi.org/10.1007/s41060-022-00311-6

Nisbet, E. C., Mortenson, C., & Li, Q. (2021). The presumed influence of election misinformation on others reduces our own satisfaction with democracy. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-59

Qiu, L. (2022, February 9). Fact-Checking Joe Rogan's Interview With Robert Malone That Caused an Uproar. *The New York Times*. https://www.nytimes.com/2022/02/08/arts/music/fact-check-joe-rogan-robert-malone.html

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1), 1–10. https://doi.org/10.1057/s41599-019-0279-9

Sprejer, L., Margetts, H., Oliveira, K., O'Sullivan, D. J. P., & Vidgen, B. (2022). An actor-based approach to understanding radical right viral tweets in the UK. Journal of Policing, Intelligence and Counter Terrorism, 18(2), 139–157. https://doi.org/10.1080/18335330.2022.2086440

Sule, S., DaCosta, M. C., DeCou, E., Gilson, C., Wallace, K., & Goff, S. L. (2023). Communication of COVID-19 Misinformation on Social Media by Physicians in the US. *JAMA Network Open*, 6(8), e2328928. https://doi.org/10.1001/jamanetworkopen.2023.28928

Sullivan, M. C. (2019). Why librarians can't fight fake news. *Journal of librarianship and information science*, 51(4), 1146-1156.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science (New York, N.Y.)*, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Winters, M., Oppenheim, B., Sengeh, P., Jalloh, M. B., Webber, N., Pratt, S. A., Leigh, B., Molsted-Alvesson, H., Zeebari, Z., Sundberg, C. J., Jalloh, M. F., & Nordenstedt, H. (2021). Debunking highly prevalent health misinformation using audio dramas delivered by WhatsApp: Evidence from a randomised controlled trial in Sierra Leone. BMJ Global Health, 6(11), e006954. https://doi.org/10.1136/bmjgh-2021-006954

Wirtschafter, V. (2023, February). Audible reckoning: How top political podcasters spread unsubstantiated and false claims. Brookings. https://www.brookings.edu/articles/audible-reckoning-how-top-political-podcasters-spread-unsubstantiated-and-false-claims/

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Beyond the hype: longitudinal trends in virtual reality perceptions beyond the COVID-19 pandemic

Christopher Ball, Kuo-Ting Huang, and Jess Francis-Levin

DOI: https://doi.org/10.47989/ir30iConf47356

Abstract

Introduction. The metaverse is shaped significantly by spatial computing technologies, such as virtual reality (VR). However, the promise of a VR-facilitated metaverse remains unfulfilled, and public perceptions of these technologies are fluid. Therefore, this study investigates perceptions and intentions to use VR over three years, both during and after the COVID-19 pandemic.

Method. We conducted three waves of cross-sectional surveys from 2020 to 2022 (N = 928), via Amazon Mechanical Turk.

Analysis. Data were analysed using OLS regression and mediation-moderation analysis using Hayes' PROCESS Model 85.

Results. Respondents' perceptions of VR as both easy to use and useful predict their intentions to use the technology, with usefulness being the stronger predictor. When examining changes over time, in 2021, VR ownership was not a predictor of perceived usefulness, and it was associated with a decrease in intentions to use VR, potentially reflecting the dynamics of the hype cycle. The direct and indirect effects of COVID-19 on VR acceptance persisted throughout all three years of the study but appear to be diminishing with time.

Conclusion. This study contributes to the theoretical and practical discourse on the metaverse's development, advocating for a nuanced understanding of VR's role as a critical component of this digital frontier.

Introduction

The metaverse concept represents a three-dimensional evolution of the internet that leverages extended reality technologies to help create an immersive, persistent, shared, and decentralized virtual world (Hwang & Chien, 2022). Virtual reality (VR), facilitated by head-mounted displays (HMDs), allows users to enter the 'metaverse' by immersing themselves in digital worlds. The potential applications of VR and the metaverse are vast and varied, as are its proponents. However, concerns and challenges persist regarding the future of this rapidly changing technology. For example, despite strides, VR adoption remains relatively low (Petrov, 2021). There are also questions about the long-term use of VR devices after adoption. For example, a report found that 69% of VR adopters use it less than once a month (AR Insider, 2020). These findings raise questions about long-term VR engagement and its viability as a gateway to the metaverse. Some scholars suggest that VR may represent a temporary trend, which demonstrates mixed results in certain applications such as advertising (Duguay, Dietzel, & Myles, 2022). Consequently, if the metaverse is to gain widespread traction, it is important to examine not only the factors contributing to VR adoption but also people's intention to use VR over time.

Concerns about VR extended beyond adoption and usage rates. Both academics and end users have expressed potential issues such as privacy and data protection within the metaverse framework (Egliston & Carter, 2022). The trajectory of VR as an integral element of the metaverse ecosystem remains an unfolding narrative. Significant industry shifts, such as Facebook's rebranding to "Meta," highlight a strategic pivot towards the metaverse, largely driven by VR technology (Meta, 2021). While this rebranding was not solely due to Meta's focus on the metaverse, its significance should not be underestimated, as Meta has been one of the most prominent investors in VR technology since it purchased Oculus in 2014 for approximately 2 billion dollars (Meta, 2014). This evolving narrative demands an in-depth exploration of public perceptions of VR, as a foundational technology of the metaverse, which are changing over time. These perceptions are not static and may significantly shape the technology's integration (or lack thereof) into everyday life moving forward.

A growing body of scholarship has examined VR acceptance and adoption across several contexts relevant to the metaverse. For example, researchers have studied the factors influencing users' willingness to adopt VR in tourism and educational contexts (Jang, Ko, Shin, & Han, 2021; Kim, Lee, & Preis, 2020). In this promising body of scholarship, the technology acceptance model (TAM) has emerged as a valuable framework for understanding VR adoption across different contexts (Ball, Huang, & Francis, 2021; Jang et al., 2021; Lee, Kim, & Choi, 2019). This body of work highlights the nuanced and evolving relationship between technology and user, particularly in the context of the metaverse. Scholars have even created a VR-specific version of TAM that adds several key variables, such as curiosity (Manis & Choi, 2019). While these studies are insightful, a common limitation of VR research in general, and VR adoption studies in particular, is the lack of longitudinal data (Ball et al., 2021), which restricts the ability to examine trends over time. This gap presents both research challenges and opportunities for further research.

Longitudinal studies are essential because attitudes toward technology may change due to historical events. For example, attitudes towards virtual worlds have been shown to vary based on time, (Luse, Mennecke, & Triplett, 2013) and device ownership status (Korucu & Bicer, 2018). Additionally, the pandemic has catalysed shifts in the digital landscape, notably impacting the perception and use of technologies like VR. Recent studies have indicated that the functionalities and affordances of technologies, such as virtual dating apps, have changed or evolved in response to the 2019 coronavirus disease (COVID-19) (Duguay et al., 2022). These shifts are critical to understanding the trajectory of VR and its integration into the metaverse post-pandemic. Moreover, studies have found that the perceived impacts of COVID-19 were related to VR perceptions at the beginning of the pandemic (Ball et al., 2021). Thus, examining VR perceptions

both during and after the COVID-19 pandemic is essential to assess if these pandemic-related effects are enduring.

Theoretical background

The technology acceptance model (TAM)

TAM has proven useful for broadly examining emerging technology adoption and VR adoption more specifically. In essence, TAM examines how potential users' perceptions that new technology is both useful and easy to use ultimately influence their intention to use (ITU) a new technology (Davis, Bagozzi, & Warshaw, 1989). Specifically, the perceived usefulness (PU) of a technology is related to a user's belief that a new technology could improve their life in a meaningful way. Furthermore, a technology's perceived ease of use (PEOU) is related to a user's belief that using a new technology will not require too much effort. Perceived ease of use can also enhance the perceived usefulness of technology, as technologies that are easier to use are often perceived as more beneficial. Both perceptions (usefulness and ease of use) subsequently impact potential users' willingness to adopt (intent to use), which is theoretically linked to technology adoption (purchase behaviour).

TAM has been used successfully for over a decade to examine VR acceptance and adoption (Fagan, Kilmon, & Pandey, 2012). This body of scholarship continues to grow as new scales are developed, and different contexts and populations are examined (Bunz, Seibert, & Hendrickse, 2021; Jang et al., 2021). As a result, scholars have built upon the TAM framework by examining key external variables that also impact VR acceptance and adoption (Castiblanco Jimenez, Cepeda García, Violante, Marcolin, & Vezzetti, 2020; Sagnier, Loup-Escande, Lourdeaux, Thouvenin, & Valléry, 2020). A notable advancement in this area is the creation of a VR-specific version of TAM, known as the 'virtual reality hardware acceptance model' (VR-HAM) (Manis & Choi, 2019). VR-HAM includes a series of external variables that form a complex model predicting VR use and purchase intentions. For instance, VR-HAM considers user variables such as curiosity and its impact on perceived ease of use. However, while the TAM literature shows promise, the results are still mixed. For example, in one study, while the perceived usefulness of VR predicted participants' intention to use VR, the perceived ease of use did not (Sagnier et al., 2020). This study also explored external factors that reduce intentions to use VR, such as cybersickness (Sagnier et al., 2020). Therefore, further research is still required to explore the factors that encourage and discourage VR perceptions and to continue testing the efficacy of TAM in a VR context.

Another crucial external user variable that has emerged in the literature is material access to the hardware or technology (Van Deursen & Van Dijk, 2019). Specifically, the digital divide literature suggests that having access to VR can influence perceptions of that technology as users gain first-hand experience, whether positive or negative (Huang, Ball, Cotten, & O'Neal, 2020). Furthermore, VR ownership does not necessarily guarantee intention to use VR, as reports have indicated that 69% of VR owners report using the hardware less than once a month (AR Insider, 2020). In other words, if people purchase VR, but their experience is negative, then they may have less intentions to use VR in the future. Thus, our study will explore how material access, or ownership of VR hardware, affects people's perceptions of VR and, ultimately, their intention to use it.

VR ownership during the COVID-19 pandemic

While user-level external variables are potentially important influencers of VR acceptance and adoption, contextual factors also play an important role. Specifically, the onset of the COVID-19 pandemic, as an external variable, presents a unique context to study VR technology adoption (Ball et al., 2021). In March 2020, most of the world instituted mandatory stay-at-home orders (i.e., lockdown procedures) to curb the spread of COVID-19 (Goodman & Schulkin, 2020). During this time, people's lives became increasingly disrupted as they are confined to their homes for extended periods. As a result, some industries experienced a 'boom' as the demand for home-based

technology and entertainment suddenly and dramatically increased. For instance, the video game industry saw a significant increase, with a 12% rise in global sales (Clement, 2021), indicating a similar potential for growth in VR usage.

The appeal of VR, including its applications in gaming and virtual travel, naturally increased as people sought new forms of entertainment and escapism during lockdowns (Ball et al., 2021). While concrete figures are challenging to obtain, some popular press articles estimate that spending on extended reality technologies increased by 50% in 2020, fuelled at least in part by the pandemic (Vardomatski, 2021). Interestingly, online surveys indicated a disparity in VR usage: while 26% of teens in the US owned a VR device, nearly half of these users seldom utilized them (Lin, 2022). This raises questions about the long-term engagement with VR technologies post-adoption.

Three years post-COVID-19, the pandemic's impact persists, though signs indicate a shift towards a more stable and predictable situation, potentially moving from a pandemic to an endemic (Grennan, 2019). By early 2023, 68.5% of the global population had been vaccinated against COVID-19 (Our World in Data, 2022), leading to a relaxation of strict containment measures. Consequently, industries that flourished during the pandemic, like video games, began to experience a downturn (Monaco-Vavrik, 2022). This fluctuating landscape underscores the need to continue examining technology adoption in a post-pandemic world.

Prior VR adoption research found that COVID-19-related factors, such as financial and health worries, are related to VR acceptance and adoption. For example, the perceived impact of COVID-19 improved the perceived usefulness of VR as the pandemic provided a potential use case for the technology (Huang, Ball, & Francis, 2023). Moreover, the pandemic influenced how people used VR, with some turning to it for work and others for coping with mental health challenges (Ball et al., 2021). In this study, we aim to extend this research by exploring the long-term impacts of COVID-19 on VR technology acceptance and adoption.

Hypotheses and research questions

To explore the influence of the pandemic on VR adoption and further our understanding of how VR is perceived over time, we have developed a conceptual model based on the Technology Acceptance Model (TAM) and its extensions. This model serves as the foundation for our hypotheses and research questions, which investigate both direct and indirect effects of VR ownership and the perceived impacts of COVID-19 on VR use (see Figure 1).

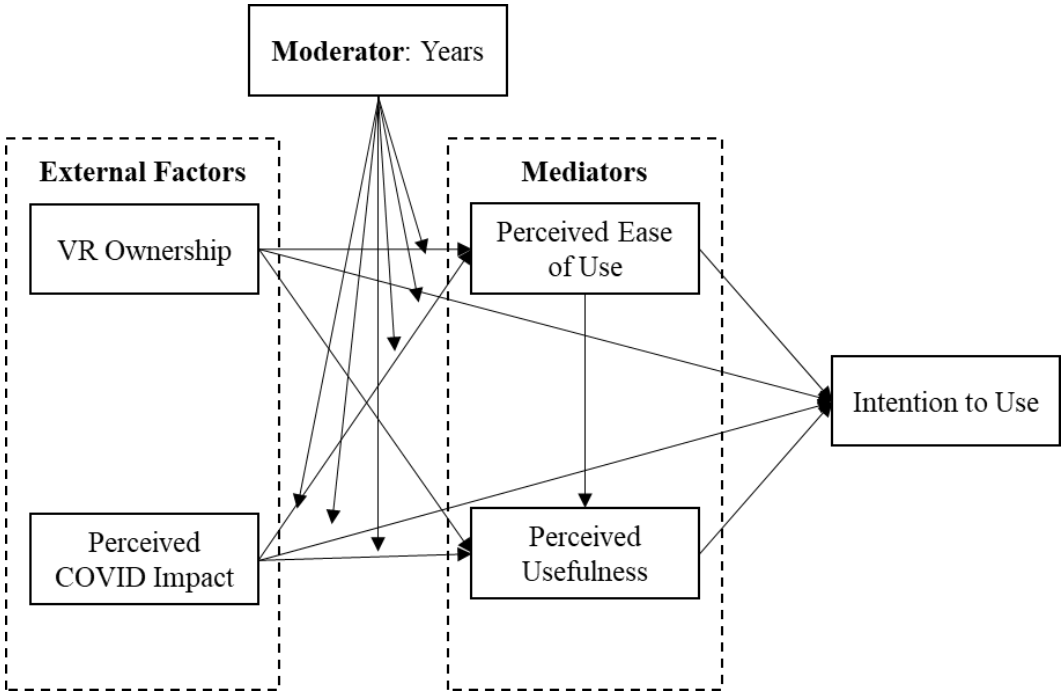


Figure 1. Proposed conceptual model

We propose the following six hypotheses based on our conceptual model:

H1: VR Ownership has a direct effect on (a) Perceived Ease of Using VR, (b) Perceived Usefulness of VR, and (c) Intention to Use VR.

H2: Perceived impacts of COVID-19 have a direct effect on (a) Perceived Ease of Using VR, (b) Perceived Usefulness of VR, and (c) Intention to Use VR.

H3: Perceived Ease of Using VR will have a positive direct effect on Perceived Usefulness of VR.

H4: (a) Perceived Ease of Using VR and (b) Perceived Usefulness of VR will have a positive direct effect on Intention to Use VR.

H5: VR Ownership has an indirect effect on Intention to Use VR via both Perceived Ease of Using VR and Perceived Usefulness of VR.

H6: Perceived impacts of COVID-19 have an indirect effect on Intention to Use VR via both Perceived Ease of Using VR and Perceived Usefulness of VR.

In addition to our hypotheses, we pose the following research questions to delve deeper into the temporal dynamics of VR adoption influenced by external variables:

RQ1: Does VR ownership impact users' PEOU, PU, and ITU related to VR technology differently in the past three years?

RQ2: Do the perceived impacts of COVID-19 impact users' PEOU, PU, and ITU related to VR technology differently in the past three years?

Methods

Participants and procedures

The current study includes a three-wave cross-sectional survey that began in 2020. All three waves of data collection were conducted through Amazon Mechanical Turk (MTurk), an online

crowdsourcing platform widely used for research purposes (Goodman et al., 2013). MTurk participants, known as 'Turkers', receive monetary compensation for completing Human Intelligence Tasks (HITs) on the MTurk platform (Paolacci & Chandler, 2014). Compensation varies based on task complexity, with Turkers earning from 1 cent to several dollars per HIT (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). The data quality from MTurk is generally considered equal to or better than data collected from undergraduate research pools (Miller, Crowe, Weiss, Maples-Keller, & Lynam, 2017).

The researchers posted three HITs on MTurk during the Fall of 2020, the Summer of 2021, and the Fall of 2022, detailing the survey purposes and associated monetary compensation. Each wave of HITs was open to participants within the United States until approximately 315 responses were received, typically over 72 hours (3 consecutive days). The target of 315 responses per wave aimed for 300 complete datasets with an additional 5% to account for potential missing or incomplete data. The sample size for each wave is in line with other studies that have examined VR adoption (Bunz et al., 2021; Jang et al., 2021; Lee et al., 2019). Respondents were not required to own a VR headset to participate in the study. A total of 928 participants were recruited for this study. Twenty-six participants with incomplete surveys or missing data were excluded, reducing the usable sample to 902. All participants received one US dollar upon submitting their survey codes to MTurk. All study procedures were approved by the Institutional Review Board at a US university, adhering to ethical standards. Respondent identifiers were removed post-incentive distribution, and results are reported in aggregate.

Survey instruments

This three-wave study uses a custom-created cross-sectional survey to investigate the impacts of COVID-19 on people's attitudes toward virtual reality. The questionnaires across all three years were composed of questions related to (1) technology acceptance and adoption (TAM), (2) COVID-19 impacts, and (3) demographics.

We utilized TAM items related to VR adoption developed by Manis and Choi (2019). These items, slightly modified from Davis's (1989) traditional TAM measures, form three scales: perceived ease of using VR (PEOU, Cronbach's $\alpha = 0.853$), perceived usefulness of VR (PU, Cronbach's $\alpha = 0.876$), and intention to use VR (ITU, Cronbach's $\alpha = 0.857$). Each scale comprises four equally weighted items, with overall scale scores ranging from 1 (lowest) to 7 (highest). Sample items include 'I believe using VR hardware would help me be more productive (PU)' and 'I believe using VR hardware would be easy for me (PEOU)'.

The COVID-19 measures employed in this study were adopted verbatim from a recent study that examined the perceived impacts of COVID-19 on psychological outcomes (Tull et al., 2020). We asked participants to report how much the pandemic impacts them in the following domains: financial worry, social support, health anxiety, and loneliness. The response categories ranged from 1(no impact at all) to 5 (impacted my life a great deal), and the items were averaged to attain an overall scale score ranging from 1 (respondent marked each of the four items as 1) to 5 (respondent marked each of the four items as 5).

We also included several VR-related and demographic control variables, including device ownership (i.e., material access) and types of devices participants owned. Lastly, we asked participants about their demographic information, including sex, age, income, education level, and race. The order of the sections presented in the questionnaire was as follows: technology acceptance variables, VR-related variables, the perceived impacts of COVID-19, and control variables.

Analytic tools

We employed descriptive analyses, ordinary least squares regressions (OLS), and mediation-moderation analyses. Descriptive analyses provided basic variable information, while OLS tested the four hypotheses. Mediation-moderation analyses, conducted using Hayes's (2018) PROCESS macro in SPSS v29, explored the two research questions regarding VR technology adoption.

Results

Descriptive analysis

The descriptive analysis included (1) demographic characteristics of participants and their VR ownership and (2) means and standardized deviations of the variables of interest. For demographic information, 61.6% of participants identified themselves as males, 69.2% were white, and more than half of the respondents were 25-34 years old. Regarding the respondents' household income, more than one-third of the participants were between $40,000 and $59,999. For the education levels, almost 80% of participants possessed a 4-year degree or higher, while less than 10% did not have some level of higher education. This participant profile is consistent with the demographic trends typically seen in the U.S. MTurk population, as noted in other studies exploring MTurk user demographics (Moss, Rosenzweig, Robinson, & Litman, 2020).

Regarding VR hardware ownership, close to 75% of respondents reported owning some form of VR hardware. This rate of VR hardware ownership in our study appears higher than the figures reported in previous VR adoption studies, such as 50% in Manis & Choi (2019) and 69.6% in Huang, Ball, and Francis (2023). However, it's noteworthy that most respondents indicated ownership of more accessible, lower-cost VR hardware, like Google Cardboard (19.7%) and Samsung Gear VR (51.2%). In contrast, only a smaller fraction, about 15%, reported possessing higher-end VR head-mounted displays (HMDs) such as the HTC Vive or the Valve Index. In other words, while VR adoption appears high in our sample (74.1%), it is essential to note that there is a great deal of variability in the type and quality of VR devices owned, and most (77%) devices our respondents reported owning are low-cost headsets with limited immersive characteristics and features such as Google Cardboard. Table 1 provides a detailed breakdown of these demographic characteristics and the specifics of VR hardware ownership among the participants.

Variable	Categories	Number of Participants (%)
Year of Data Collected	2020	312 (33.6%)
	2021	303 (32.7%)
	2022	313 (33.7%)
Sex	Male	556 (61.6%)
	Female	344 (38.1%)
	Non-Binary/Third Gender	2 (0.2%)
Age	18-24	65 (7.2%)
	25-34	465 (51.6%)
	35-44	208 (23.1%)
	45-54	108 (12.0%)
	55-64	42 (4.7%)
	65 and above	14 (1.6%)
Household Income	Less than $10,000	41 (4.6%)
	$10,000-$19,999	63 (7.0%)
	$20,000-$29,999	88 (9.8%)
	$30,000-$39,999	93 (10.3%)
	$40,000-$49,999	146 (16.2%)
	$50,000-$59,999	162 (18.0%)
	$60,000-$69,999	74 (8.2%)
	$70,000-$79,999	82 (9.1%)
	$80,000-$89,999	53 (5.9%)
	$90,000-$99,999	49 (5.4%)
	$100,000-$149,999	41 (4.6%)
	More than $150,000	9 (1.0%)
Education Level	Less than high school	2 (0.2%)
	High school graduate	27 (3.0%)
	Some college	53 (5.9%)
	2-year degree	28 (3.1%)
	4-year degree	617 (68.9%)
	Master's degree	161 (18.0%)
	Doctorate	8 (0.9%)
Race/Ethnicity	White	621 (69.2%)
	Black/African American	107 (11.9%)
	American Indian or Alaska Native	16 (1.8%)
	Asian	135 (15.0%)
	Others	19 (2.1%)
VR Hardware Ownership	No	234 (25.9%)
	Yes	668 (74.1%)
Types of VR Hardware (Check all that apply)	Google Cardboard	178 (19.7%)
	Samsung Gear VR	462 (51.2%)
	Oculus Go	55 (6.1%)
	Oculus Rift(s)	57 (6.3%)
	Oculus Quest	78 (8.6%)
	HTC Vive	53 (5.9%)
	PlayStation VR	132 (14.6%)
	Valve Index	38 (4.2%)

Table 1. The demographic characteristics of the participants (N=902)

The variables of interest include VR hardware ownership, the perceived impacts of COVID-19, and TAM variables. Almost 75% of respondents reported owning VR hardware, although most own entry-level devices such as Samsung Gear VR and Google Cardboard. For the perceived impacts of COVID-19, the means were calculated based on the responses of the 902 participants. Participants reported moderate overall perceived impact of COVID-19 (Cronbach's $\alpha = 0.853$). Participants also reported a moderate amount of COVID-19 impacts across all four categories. Financial worry and lack of social support are the top two perceived impacts, followed by loneliness and health anxiety.

Regarding TAM-related variables, participants reported an average of 5.47 and 5.61 points out of a 7-point Likert-like scale (7 as the highest) when asked about the PEOU and PU of VR hardware. Furthermore, participants also reported a strong intention to use (5.43) on average. Table 2 presents the descriptive analysis of the perceived impacts of COVID-19 and the TAM variables.

Construct (Numbers of Items; Cronbach's α)	Response Categories	Mean (Standard Deviation)
Perceived Impacts of COVID-19 (0.853)	1= None	3.79 (1.32)
Health Anxiety	2= A little	3.68 (1.50)
Financial Worry	3= A moderate amount	3.86 (1.60)
Lack of Social Support	4 = A lot	3.81 (1.56)
Loneliness	5= A great deal	3.79 (1.68)
Perceived Usefulness (PU) (4 items; 0.876) Example Item: Using VR hardware would be useful in my life.	From 1 to 7 1= Strongly disagree 7=Strongly agree	5.47 (1.07)
Perceived Ease of Use (PEOU) (4 items; 0.839) Example Item: I believe using VR hardware would be easy for me.	From 1 to 7 1= Strongly disagree 7=Strongly agree	5.61 (0.89)
Intention to Use (ITU) (4 items; 0.857) Example Item: I will use VR hardware within the foreseeable future.	From 1 to 7 1= Strongly disagree 7=Strongly agree	5.43 (1.10)

Table 2. Means and standard deviations of the variables of interest

OLS regression results

A series of OLS regressions were conducted to test our hypotheses, with results detailed in Table 3. In Model 1, the focus was on the direct effects of VR ownership and the perceived impacts of COVID-19 on participants' perceived ease of using VR (PEOU), addressing Hypotheses H1 and H2. The main predictors were VR ownership and the perceived impacts of COVID-19, with demographic variables such as sex, age, race, educational level, and household income as control variables. Model 2 explored the relationship between PEOU and perceived usefulness (PU), in line with Hypothesis H4. Here, PU was the dependent variable, with PEOU included as a critical predictor along with the variables from Model 1.

Model 3 examined the direct effects of VR ownership and COVID-19 on Intention to Use VR (ITU), relevant to Hypotheses H1 and H2. This model integrated the main variables - VR ownership and COVID-19 impacts - and demographic variables as predictors. Model 4 expanded the analysis to include PEOU and PU as predictors of ITU, crucial for exploring the potential mediating relationships posited in Hypotheses H3, H5, and H6. It compared the differences between Models 3 and 4 to assess mediation. There is no multicollinearity (VIF < 4) in all four models, and the values of skewness (between -2 to +2) and kurtosis (between-7 to +7) were considered acceptable (Hair et al., 2010).

The OLS regression results from Models 1-4 provided insightful findings. In Models 1, 2, and 3, the analyses revealed that VR ownership significantly predicted perceived usefulness (PU, β =.12, p < .001) and intention to use VR (ITU, β =.16, p < .001) but did not significantly predict perceived ease of use (PEOU). This *partial support for H1* suggests a complex relationship between VR ownership and its perceived attributes. On the other hand, the perceived impacts of COVID-19 significantly influenced PEOU (β =.26, p < .001), PU (β =.15, p < .001), and ITU (β =.33, p < .001), offering *full support for H2* and highlighting the significant role of the pandemic in shaping attitudes towards VR. The findings also showed a positive influence of PEOU on PU (β =.65, p < 0.001), *fully supporting H3*. This underscores the importance of ease of use in enhancing the perceived usefulness of VR.

In Models 3 and 4, mediation relationships were further explored. Both PEOU (β =.28, p < 0.001) and PU (β =.59, p < 0.01) emerged as significant predictors of ITU in Model 4, *fully supporting H4*, with PU being the more influential factor. Notably, the effect of VR ownership on ITU was mediated by PEOU and PU, as evidenced by a more than 56% decrease in its effect (from .16 in Model 3 to .07 in Model 4), *supporting H5*. Similarly, the impact of COVID-19 on ITU also showed a significant decrease (almost 79% from .33 in Model 3 to .07 in Model 4), indicating an indirect effect via PEOU and PU and *supporting H6*.

Demographic variables played a notable role as well. Household income (β =.07, p < 0.05) was a predictor of PEOU, and education level (β =.08, p < 0.001) influenced PU. These findings emphasize the importance of demographic factors in the adoption and perception of VR technology.

	Model 1		Model 2		Model 3		Model 4	
Dependent Variables	**Ease of Use**		**Usefulness**		**Use Intention**		**Use Intention**	
Sex (1=male, 0 = female)	.02		-.01		.01		.01	
Age	.01		-.03		.01		.02	
Race (1= white, 0=non-white)	.01		-.04		.00		-.02	
Household Income	.07	*	-.02		.07	*	.03	
Education Level	.00		.08	**	.08	*	.04	
VR ownership	.03		.12	***	.16	***	.07	***
COVID-19	.26	***	.15	***	.33	***	.07	***
Ease of Use			.65	***			.28	***
Usefulness							.59	***
F	10.39	***	126.78	***			291.92	***
Adjusted R^2	.07		.53				.75	

Table 3. OLS regression analysis of technology acceptance variables regressed on perceived impact of COVID-19 and demographics

Mediation-moderation results

To answer our research questions, we conducted two mediation-moderation analyses using PROCESS model 85 with 5,000 bootstrap samples and a 95% confidence interval. In the first PROCESS analysis, we explored VR ownership as the independent variable and ITU as the dependent variable. The year of data collection was a moderator, and PEOU and PU served as mediators. Demographic variables and COVID-19 were included as covariates.

Our analysis revealed significant interaction effects of VR ownership on PU (F (2,872) =7.5594, p<.001) and ITU (F (2,872) =10.6711, p<.001). However, there were no interaction effects of VR ownership on PEOU. Regarding conditional direct effects, VR ownership positively impacted PU in 2020 (β = .2863, p < .01) and 2022 (β = .3721, p < .001), but not statistically significant in 2021 (β = -.1664, p = .117). The direct effect on ITU was negatively significant in 2021 (β = -.3546, p < .001), with non-significant positive effects in 2020 and 2022.

The conditional indirect effects through PEOU and PU varied annually. The indirect effect through PEOU was not significant across the three years. Through PU, the effect was significant in 2020 (β = .1755, 95% CI [.0305, .3518]) and 2022 (β = .2281, 95% CI [.0670, .4114]), but not significant in 2021. The combined mediators' effect was not significant across the three-year span.

These findings suggest a complex and fluctuating relationship between VR ownership and user perceptions across different years, especially highlighted by the contrasting results in 2021. Table 4 details VR ownership's conditional direct and indirect effects on the outcome variables.

Conditional Effect of VR Ownership on Perceived Ease of Use (PEOU)					
		Effect	Se	T	p
VR ownership ➔	PEOU				
	Year 2020	N/A			
	Year 2021	N/A			
	Year 2022	N/A			
Conditional Effect of VR Ownership on Perceived Usefulness (PU)					
		Effect	Se	T	p
VR ownership ➔	PU				
	Year 2020	**.2863**	**.0929**	**3.0810**	**.002*****
	Year 2021	-.1664	.1060	-1.5700	.117
	Year 2022	**.3721**	**.1006**	**3.6989**	**.000*****
Conditional Effect of VR Ownership on Intention to Use (ITU)					
		Effect	Se	T	P
VR ownership ➔	ITU				
	Year 2020	.1190	.0711	1.6742	.094
	Year 2021	**-.3546**	**.0807**	**-4.3913**	**.000*****
	Year 2022	.0730	.0771	.9458	.345
Conditional Indirect Effect of VR Ownership on Intention to Use (ITU)					
		Effect	BootSE	BootLLCI	BootULCI
VR ownership ➔	PEOU ➔ ITU				
	Year 2020	.0287	.0405	-.0462	.1122
	Year 2021	-.0677	.0610	-.1976	.0421
	Year 2022	-.0532	.0460	-.1452	.0370
		Index	BootSE	BootLLCI	BootULCI
	Year 2021	-.0965	.0750	-.2570	.0396
	Year 2022	-.0819	.0627	-.2143	.0364
		Effect	BootSE	BootLLCI	BootULCI
VR ownership ➔	PU ➔ ITU				
	Year 2020	**.1755**	**.0818**	**.0305**	**.3518**
	Year 2021	-.1020	.0821	-.2693	.0536
	Year 2022	**.2281**	**.0881**	**.0670**	**.4114**
		Index	BootSE	BootLLCI	BootULCI
	Year 2021	-.2774	.1197	-.5256	-.0586
	Year 2022	.0526	.1166	.1755	.2895
		Effect	BootSE	BootLLCI	BootULCI
VR ownership ➔	PEOU ➔ PU ➔ ITU				
	Year 2020	.0396	.0543	-.0629	.1495
	Year 2021	-.0933	.0790	-.2452	.0602
	Year 2022	-.0733	.0611	-.1927	.0495
		Index	BootSE	BootLLCI	BootULCI
	Year 2021	-.1329	.0966	-.3243	.0560
	Year 2022	-.1129	.0832	-.2807	.0520

Table 4. Conditional direct and indirect effects of VR ownership on TAM variables

In the second PROCESS analysis, the impact of COVID-19 was examined as the independent variable, with intention to use (ITU) VR as the dependent variable. The year of data collection was utilized as a moderator, while perceived ease of use (PEOU) and perceived usefulness (PU) were mediators. Additional covariates included demographic variables and VR ownership.

Our analysis indicated significant interaction effects of the COVID-19 impact on PU ($F(2,872) = 7.5594$, $p < .001$) and ITU ($F(2,872) = 10.6711$, $p < .001$), but not PEOU. Specifically, the conditional direct effects revealed a varying influence of the pandemic on users' perceptions and intentions to use VR across different years. In 2020, the impact of COVID-19 positively influenced PU ($\beta = .2986$, $p < .001$), suggesting an increased perception of VR's utility amidst the early pandemic. However,

in 2021, this effect was not as pronounced, indicating a possible adjustment in perceptions as the pandemic progressed (β = .0958, p < .001). By 2022, the positive impact on PU re-emerged strongly (β = .3140, p < .001), reflecting a sustained or growing appreciation of VR's usefulness as the pandemic evolved.

Regarding the indirect effects of COVID-19 on users' intention to use VR (ITU) via perceived ease of use (PEOU) and perceived usefulness (PU), the results are different across three years. In 2020, the impact of COVID-19 on ITU through PEOU was notable (β = .1021, 95% CI [.0579, .1556]), highlighting the initial phase of the pandemic where the perception of VR's ease of use significantly influenced usage intentions. This pathway remained significant in 2021 (β = .0752, 95% CI [.0429, .1173]) and persisted in 2022 (β = .0654, 95% CI [.0234, .1194]), underscoring the sustained influence of ease-of-use perceptions on VR usage intentions as the pandemic changed.

Similarly, the indirect impact of COVID-19 on ITU via PU was significant across each year, with the effect being considerable in 2020 (β = .1197, 95% CI [.0392, .2037]), continuing in 2021 (β = .0592, 95% CI [.0161, .1094]), and becoming particularly strong in 2022 (β = .1941, 95% CI [.0995, .3046]). This trend indicates a growing recognition of VR's usefulness throughout the pandemic's progression.

Additionally, the combined effect of PEOU and PU as mediators was significant. In 2020, their combined influence on the relationship between COVID-19 and ITU was substantial (β = .1413, 95% CI [.0825, .2045]). This effect was again significant in 2021 (β = .1040, 95% CI [.0627, .1492]) and continued into 2022 (β = .0905, 95% CI [.0334, .1577]), highlighting the enduring, but potentially waning, impact of the pandemic on VR technology acceptance. See Table 5 below for the conditional direct and indirect effects of COVID-19 on TAM variables.

Conditional Effect of COVID-19 on Perceived Ease of Use (PEOU)					
		Effect	Se	t	p
COVID-19 ➔	PEOU				
	Year 2020	N/A			
	Year 2021	N/A			
	Year 2022	N/A			
Conditional Effect of COVID-19 on Perceived Usefulness (PU)					
		Effect	Se	t	p
COVID-19 ➔	PU				
	Year 2020	**.1937**	**.0524**	**3.6946**	**.000*** **
	Year 2021	**.0958**	**.0289**	**3.3141**	**.000*** **
	Year 2022	**.3140**	**.0475**	**6.6087**	**.000*** **
Conditional Effect of COVID-19 on Intention to Use (ITU)					
		Effect	Se	t	p
COVID-19 ➔	ITU				
	Year 2020	.0170	.0405	.4184	.676
	Year 2021	**.1132**	**.0223**	**5.0713**	**.000*** **
	Year 2022	.0237	.0374	.6336	.527
Conditional Indirect Effect of COVID-19 on Intention to Use (ITU)					
		Effect	BootSE	BootLLCI	BootULCI
COVID-19 ➔	PEOU ➔ ITU				
	Year 2020	**.1021**	**.0251**	**.0579**	**.1556**
	Year 2021	**.0752**	**.0192**	**.0429**	**.1173**
	Year 2022	**.0654**	**.0243**	**.0234**	**.1194**
		Index	BootSE	BootLLCI	BootULCI
	Year 2021	-.0270	.0255	-.0786	.0233
	Year 2022	-.0367	.0305	-.0988	.0227
		Effect	BootSE	BootLLCI	BootULCI
COVID-19 ➔	PU ➔ ITU				
	Year 2020	**.1197**	**.0428**	**.0392**	**.2037**
	Year 2021	**.0592**	**.0238**	**.0161**	**.1094**
	Year 2022	**.1941**	**.0525**	**.0995**	**.3046**
		Index	BootSE	BootLLCI	BootULCI
	Year 2021	-.0605	.046	-.1501	.0313
	Year 2022	.0744	.0636	-.0452	.2056
		Effect	BootSE	BootLLCI	BootULCI
COVID-19 ➔	PEOU ➔ PU ➔ ITU				
	Year 2020	**.1413**	**.0308**	**.0825**	**.2045**
	Year 2021	**.1040**	**.0223**	**.0627**	**.1492**
	Year 2022	**.0905**	**.0320**	**.0334**	**.1577**
		Index	BootSE	BootLLCI	BootULCI
	Year 2021	-.0373	.0348	-.1054	.0314
	Year 2022	-.0508	.0411	-.1292	.0309

Table 5. Conditional direct and indirect effects of COVID-19 on TAM variables

These findings demonstrate a complex interplay where both ease of use and usefulness perceptions significantly mediate the relationship between the pandemic and the intention to use VR, emphasizing how external events like COVID-19 can shape technology adoption over time. Table 6 summarizes the findings of the hypotheses and research questions.

Hs/RQs	Description	Findings
H1a	VR Ownership to PEOU	Not supported
H1b	VR Ownership to PU	Supported
H1c	VR Ownership to ITU	Supported
H2a	COVID-19 to PEOU	Supported
H2b	COVID-19 to PU	Supported
H2c	COVID-19 to ITU	Supported
H3	PEOU to PU	Supported
H4a	PEOU to ITU	Supported
H4b	PU to ITU	Supported
H5	Indirect effect of VR Ownership on ITU via PEOU and PU	Supported
H6	Indirect effect of COVID-19 on ITU via PEOU and PU	Supported
RQ1	Moderation by Year on VR Ownership's direct and indirect effect on TAM variables	Yes, notably different in 2021
RQ2	Moderation by Year on Perceived COVID-19's direct and indirect impact on TAM variables	Yes, strongest in 2020 & 2022 for PU and ITU

Table 6. Findings of hypotheses and research questions

Discussion

Summary and interpretation of findings

The study offers valuable insights into the acceptance and adoption of VR technologies, particularly within the evolving context of the metaverse. Our results confirm the role of perceived ease of use (PEOU) and perceived usefulness (PU) in shaping users' intentions to use VR (ITU), which is consistent with existing literature (Kim et al., 2020; Lee et al., 2019; Manis & Choi, 2019). However, the results also reveal more nuanced relationships. Specifically, while VR ownership predicts PU and ITU, it does not significantly influence PEOU. This finding challenges the assumption that VR ownership inherently translates into improved usability perception, which suggests that while owners recognize the utility of VR technology, usability remains a persistent barrier.

This study further showed the significant role of COVID-19 as an external factor that positively influenced PEOU, PU, and ITU. The perceived impacts of the pandemic demonstrate its role in accelerating VR adoption and reinforcing its role as a gateway technology to the metaverse. Our mediation analysis further revealed that the impact of VR ownership and COVID-19 on ITU is indirect, which is mediated by PU and PEOU. Notably, the direct effect of VR ownership on ITU was significantly reduced when considering these mediating factors, indicating the importance of both ease and usefulness perceptions in shaping adoption behaviours.

Our focus on the temporal dynamics of VR technology acceptance revealed interesting patterns. The impact of VR ownership on TAM variables varied across the years, with 2021 presenting a unique scenario where VR ownership negatively correlated with intention to use VR devices. This could indicate a shift in user expectations during this period. Similarly, the indirect effects of COVID-19 on ITU via PEOU and PU were consistent across all three years, but the effects showed a gradual decline, suggesting an enduring, yet slowly diminishing, influence of the pandemic on VR acceptance.

In summary, our findings highlight a complex interplay between user perceptions, external influences, and temporal dynamics in shaping the adoption and acceptance of VR technologies (See Figure 2). While VR ownership drives perceived usefulness, usability concerns remain unresolved, signalling a need for enhanced user experiences. The pandemic's role as a catalyst for VR adoption remains evident but is slowly diminishing. These insights are particularly relevant in understanding VR's evolving role as a gateway to the metaverse and provide a foundational basis for future research in this rapidly changing field.

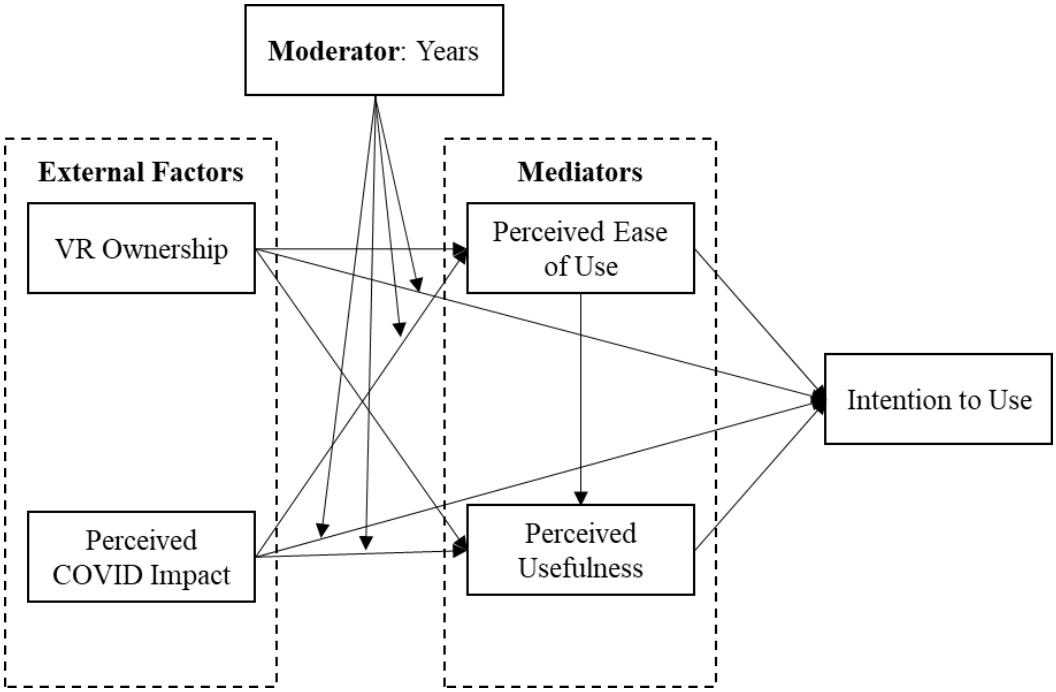


Figure 2. Revised conceptual model

Theoretical contributions

The current study has three principal theoretical contributions. First, our study significantly contributes to the theoretical understanding of VR adoption within the technology acceptance model (TAM) framework, particularly in the context of the emerging metaverse. The findings corroborate prior research (Kim et al., 2020; Lee et al., 2019; Manis & Choi, 2019), emphasizing the vital influence of perceived ease of use and usefulness in determining users' intentions to use VR technology. This reiteration of TAM's foundational components in a novel context in which public awareness of the metaverse has flowed and ebbed, underlines the model's robustness and adaptability in explaining technology acceptance across evolving digital landscapes.

Second, our longitudinal approach provides insights into the evolution of VR acceptance, especially during and after the COVID-19 pandemic. This period, marked by substantial shifts in user behaviour and technology interaction, is a crucial backdrop for understanding the dynamic nature of VR acceptance. We extend traditional TAM models by integrating external factors, such as the ramifications of COVID-19, which aligns with recent studies investigating health anxiety and its impact on VR adoption (Ball et al., 2021; Huang et al., 2023). This extension reflects a broader scholarly trend of incorporating external environmental and psychological factors into TAM, offering a more holistic view of technology acceptance dynamics.

Third, our findings also shed light on the nuanced relationship between VR ownership and user perceptions over time. The distinct pattern observed in 2021, where VR ownership correlated with decreased intentions to use VR, challenges and enriches the existing literature. This deviation can be theoretically framed through the concept of hype cycles (Dedehayir & Steinert, 2016; Grundmeyer, 2014; Prinsloo & Van Deventer, 2017). The apparent cycle of heightened acceptance in 2020, followed by disillusionment in 2021 and recovery in 2022, illustrates the complex interplay between technology adoption, societal trends, and marketing dynamics. This nuanced understanding of VR acceptance, influenced by hype dynamics, not only adds depth to TAM's applicability in new technological domains but also underscores the need for a temporal and contextual lens in technology adoption research.

Practical implications

Our findings offer vital insights for VR technology developers and manufacturers, especially in the context of the evolving metaverse. The fluctuating perceptions of VR's usefulness and ease of use, particularly noted in 2021, highlight the need for continuous innovation in user interfaces and overall user experience. Developers should focus on enhancing accessibility and providing robust support to address potential user frustrations and facilitate long-term engagement with VR technologies. Marketing professionals can leverage these insights to refine their strategies, focusing on the practicality and versatility of VR applications. During periods of decreased consumer interest, as observed in our study, emphasizing VR's diverse applications can help sustain user engagement and drive adoption.

The past decade has seen notable advancements in making VR more accessible, with a significant reduction in the price of high-quality VR devices and the simplification of setup processes. Our research confirms the positive impact of these advancements on users' intention to use VR technology. However, it also highlights the crucial role of usefulness as a consistent predictor of VR acceptance. Thus, while focusing on ease of use is vital, VR developers and designers should not overlook the importance of ensuring that VR applications are perceived as genuinely useful. Developing a "killer app" that showcases a unique and compelling use of VR could be a key driver for broader adoption and acceptance of VR in the mainstream market.

The consistent relevance of VR throughout the pandemic period signals a stable interest in immersive spatial computing technologies, presenting lucrative opportunities for investors and businesses. Strategic investments to bridge the gap between virtual and augmented realities could lead to significant advancements, contributing to a more immersive and cohesive metaverse experience.

Limitations and future directions

Our study contributes significantly to the understanding of VR adoption, yet it has limitations that need addressing in future research. A primary limitation lies in not accounting for the variability of VR devices used by participants. The type and quality of VR devices can greatly impact user experiences and perceptions. Hence, future studies should differentiate between various VR devices to provide a more nuanced understanding of VR adoption.

Second, relying on an MTurk panel sample may not fully represent the general population. MTurk samples tend to overrepresent younger individuals, potentially skewing results (Fleischer, Mead, & Huang, 2015; Huff & Tingley, 2015). Furthermore, this was a panel study, so participants varied over time. Future research should thus employ different methods, such as a cohort panel design or more diverse sampling methods, like probability sampling, to enhance the generalizability of the findings.

Third, to explore the impact of hardware access on the intention to use VR, we examined user's VR ownership. However, other forms of access, such as borrowing a headset, may play a significant role. Future studies may wish to examine the impact of broader forms of access. Lastly, our study's focus on U.S. residents limits its global applicability. Given the diverse ways VR is used and perceived worldwide, future studies should explore VR acceptance across different cultures and geographical contexts to gain a more global perspective.

Conclusion

We find ourselves on the precipice of a potential paradigm shift in how we engage with the digital world. Our findings reveal a nuanced interplay between VR ownership, perceived utility, and ease of use, with variances across years. This shift indicates a broader evolution in user attitudes, transitioning from traditional applications of VR to envisioning it as an integral component of an immersive, three-dimensional internet experience. Our findings also highlight the need for

continuous innovation and adaptation in VR technology to meet users' changing demands and expectations in the metaverse era.

Challenges and opportunities mark the journey of VR as we enter the metaverse. Our study emphasizes the critical role of perceived usefulness and ease of use in VR adoption, with usefulness emerging as the more influential factor. This insight is crucial for VR developers and stakeholders, who must focus not only on making VR accessible but also on enhancing its practical applications within the metaverse. As we enter the metaverse via spatial computing technologies, such as VR, understanding and addressing user-centric factors will be vital to ensuring the success of these technologies.

In conclusion, our research provides a valuable foundation for understanding VR's evolving role as a gateway to the metaverse. Positioned at the nexus of technological innovation and user experience, the insights gleaned from this study will be instrumental in guiding the future development and adoption of VR in the ever-expanding metaverse.

Acknowledgements

The authors would like to acknowledge that effort on this publication is supported, in part, by the National Institutes of Health under the award number K01AG083120. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

About the authors

Christopher Ball is an Assistant Professor in the Department of Journalism and Institute of Communications Research, University of Illinois at Urbana-Champaign, Illinois, USA. He received his Ph.D. from Michigan State University. His research interests include studying interactive and immersive media technologies and how they can be leveraged to address social problems such as digital inequality. He can be contacted at drball@illinois.edu

Kuo-Ting (Tim) Huang is an Assistant Professor in the School of Computing and Information, University of Pittsburgh, Pennsylvania, USA. He received his Ph.D. from Michigan State University. His research investigates the potential of emerging technologies to bridge gaps in health and education for underserved and disadvantaged populations. He can be contacted at timhuang@pitt.edu

Jess Francis-Levin is a Research Investigator at the Institute for Social Research, University of Michigan, Michigan, USA. She received her Ph.D. from Michigan State University. Her research focuses on the ability of emerging technologies to benefit individuals at risk for social isolation, especially older adult, marginalized, and rural populations. She can be contacted at jessfran@umich.edu

References

AR Insider. (2020). New Report: 19% of U.S. Adults Have Tried VR. Retrieved from https://arinsider.co/2020/05/20/new-report-19-of-u-s-adults-have-tried-vr/

Ball, C., Huang, K.-T., & Francis, J. (2021). Virtual reality adoption during the COVID-19 pandemic: A uses and gratifications perspective. *Telematics and Informatics*, 65, 101728.

Bunz, U., Seibert, J., & Hendrickse, J. (2021). From TAM to AVRTS: development and validation of the attitudes toward Virtual Reality Technology Scale. *Virtual Reality*, 25(1), 31-41.

Castiblanco Jimenez, I. A., Cepeda García, L. C., Violante, M. G., Marcolin, F., & Vezzetti, E. (2020). Commonly used external TAM variables in e-learning, agriculture, and virtual reality applications. *Future Internet*, *13*(1), 7.

Clement, J. (2021). COVID-19 impact on digital games revenue 2020, by type. Retrieved from https://www.statista.com/statistics/190184/covid-19-impact-on-digital-games-revenue-type/

Davis, F. D. (1985). A *technology acceptance model for empirically testing new end-user information systems: Theory and results*. Massachusetts Institute of Technology,

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management science*, *35*(8), 982-1003.

Dedehayir, O., & Steinert, M. (2016). The hype cycle model: A review and future directions. *Technological Forecasting and Social Change*, *108*, 28-41.

Duguay, S., Dietzel, C., & Myles, D. (2022). The year of the "virtual date": Reimagining dating app affordances during the COVID-19 pandemic. *New Media & Society*, 14614448211072257.

Egliston, B., & Carter, M. (2022). Oculus imaginaries: The promises and perils of Facebook's virtual reality. *New Media & Society*, *24*(1), 70-89.

Fagan, M., Kilmon, C., & Pandey, V. (2012). Exploring the adoption of a virtual reality simulation: The role of perceived ease of use, perceived usefulness, and personal innovativeness. *Campus-Wide Information Systems*.

Fishbein, M., & Ajzen, I. (1977). Belief, attitude, intention, and behaviour: An introduction to theory and research.

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in MTurk and other online samples. *Industrial and Organizational Psychology*, *8*(2), 196-202.

Goodman, R., & Schulkin, D. (2020). Timeline of the coronavirus pandemic and US response. In: May.

Grennan, D. (2019). What is a Pandemic? *Jama*, *321*(9), 910-910.

Grundmeyer, T. (2014). Adopting Technology: Using Student Qualitative Data and Gartner's Hype Cycle. *Journal of Education and Training Studies*, *2*(1), 207-216.

Hayes, A. F. (2018). Introduction to mediation, moderation, and conditional process analysis second edition: A regression-based approach.: New York, NY. In: Guilford Publications, Inc.

Huang, K.-T., Ball, C., Cotten, S. R., & O'Neal, L. (2020). Effective experiences: A social cognitive analysis of young students' technology self-efficacy and STEM attitudes. *Social Inclusion*, *8*(2), 213-221.

Huang, K.-T., Ball, C., & Francis, J. (2023). The Perceived Impacts of COVID-19 on Users' Acceptance of Virtual Reality Hardware: A Digital Divide Perspective. *American Behavioural Scientist*, 00027642231156775.

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 2053168015604648.

Hwang, G.-J., & Chien, S.-Y. (2022). Definition, roles, and potential research issues of the metaverse in education: An artificial intelligence perspective. *Computers and Education: Artificial Intelligence*, 100082.

Jang, J., Ko, Y., Shin, W. S., & Han, I. (2021). Augmented reality and virtual reality for learning: An examination using an extended technology acceptance model. *IEEE access*, 9, 6798-6809.

Kim, M. J., Lee, C.-K., & Preis, M. W. (2020). The impact of innovation and gratification on authentic experience, subjective well-being, and behavioural intention in tourism virtual reality: The moderating role of technology readiness. *Telematics and Informatics*, 49, 101349.

Korucu, A. T., & Bicer, H. (2018). Investigation of post-graduate Students' attitudes towards Mobile learning and opinions on mobile learning. *International Technology and Education Journal*, *2*(1), 21-34.

Lee, J., Kim, J., & Choi, J. Y. (2019). The adoption of virtual reality devices: The technology acceptance model integrating enjoyment, social interaction, and strength of the social ties. *Telematics and Informatics*, 39, 37-48.

Lin, C. (2022). If the metaverse is the future of social media, teens aren't convinced. Retrieved from https://www.fastcompany.com/90740073/if-the-metaverse-is-the-future-of-social-media-teens-arent-convinced

Luse, A., Mennecke, B., & Triplett, J. (2013). The changing nature of user attitudes toward virtual world technology: A longitudinal study. *Computers in Human Behavior*, *29*(3), 1122-1132.

Manis, K. T., & Choi, D. (2019). The virtual reality hardware acceptance model (VR-HAM): Extending and individuating the technology acceptance model (TAM) for virtual reality hardware. *Journal of Business Research*, 100, 503-513.

Meta. (2014). Facebook to Acquire Oculus. Retrieved from https://about.fb.com/news/2014/03/facebook-to-acquire-oculus/

Meta. (2021). Introducing Meta: A Social Technology Company. Retrieved from https://about.fb.com/news/2021/10/facebook-company-is-now-meta/

Miller, J. D., Crowe, M., Weiss, B., Maples-Keller, J. L., & Lynam, D. R. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. *Personality Disorders: Theory, Research, and Treatment*, 8(1), 26.

Monaco-Vavrik. (2022). Video Game Sales Set to Fall This Year After Pandemic Boom, Predicts Ampere Analysis. Retrieved from https://www.cnet.com/tech/gaming/video-game-sales-set-to-fall-this-year-after-pandemic-boom-predicts-ampere-analysis/

Moss, A. J., Rosenzweig, C., Robinson, J., & Litman, L. (2020). Is it Ethical to Use Mechanical Turk for Behavioural Research? Relevant Data from a Representative Survey of MTurk Participants and Wages.

Our World in Data. (2022). Coronavirus (COVID-19) Vaccinations. Retrieved from https://ourworldindata.org/covid-vaccinations

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184-188.

Petrov, C. (2021, March 18, 2021). 43 Virtual Reality Statistics That Will Rock the Market in 2020. Retrieved from https://techjury.net/blog/virtual-reality-statistics/#gref

Prinsloo, T., & Van Deventer, J. P. (2017). *Using the Gartner Hype Cycle to evaluate the adoption of emerging technology trends in higher education–2013 to 2016.* Paper presented at the Emerging Technologies for Education: Second International Symposium, SETE 2017, Held in Conjunction with ICWL 2017, Cape Town, South Africa, September 20–22, 2017, Revised Selected Papers 2.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI'10 extended abstracts on Human factors in computing systems* (pp. 2863-2872).

Sagnier, C., Loup-Escande, E., Lourdeaux, D., Thouvenin, I., & Valléry, G. (2020). User acceptance of virtual reality: an extended technology acceptance model. *International Journal of Human–Computer Interaction*, *36*(11), 993-1007.

Tull, M. T., Edmonds, K. A., Scamaldo, K., Richmond, J. R., Rose, J. P., & Gratz, K. L. (2020). Psychological Outcomes Associated with Stay-at-Home Orders and the Perceived Impact of COVID-19 on Daily Life. *Psychiatry research*, 113098.

Van Deursen, A. J., & Van Dijk, J. A. (2019). The first-level digital divide shifts from inequalities in physical access to inequalities in material access. *new media & society*, *21*(2), 354-375.

Vardomatski, S. (2021). Augmented And Virtual Reality After Covid-19. Retrieved from https://www.forbes.com/sites/forbestechcouncil/2021/09/14/augmented-and-virtual-reality-after-covid-19/?sh=79161f302d97

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Engaging with information beyond vision: hands-on approaches to computational thinking for blind and visually impaired learners

Kyungwon Koh, JooYoung Seo, Si Chen, and Eugene Malcolm Cox
DOI: https://doi.org/10.47989/ir30iConf47353

Abstract

Introduction. This project develops accessible maker tools and activities to foster computational thinking (CT) skills in blind and visually impaired (BVI) learners, while investigating the experiences of two key groups: (1) BVI learners and (2) librarians and maker professionals who design and deliver accessible CT programs.

Methods. The pilot phase designed and delivered an accessible electronics and coding curriculum to three BVI youth in a two-day summer camp. Data was collected through two debrief focus groups—one with BVI learners and one with the maker professionals who served as instructors.

Analysis. All interviews were recorded and transcribed. The research team used a grounded theory approach to analyse the interview data.

Results. Both learners and instructors highlighted the benefits of tactile and multi-sensory learning tools, though challenges emerged with the text-based coding platform. Learners self-reported increased confidence, autonomy, and interest in CT skills. Instructors adapted their approaches with detailed verbal descriptions and modifications to tools and lesson plans. Understanding the diverse needs of BVI learners and providing personalized assistance was crucial.

Conclusion. Tactile and physical approaches to computational thinking show promise for previously marginalized learners, though challenges remain. Future research will explore how emerging technologies, including AI, can further enhance accessibility for BVI learners.

Introduction

In contemporary culture, information is predominantly received and produced visually, whether through text, graphics, or videos. Learning resources and environments, including those in STEM fields, are no exception, which often marginalizes individuals who are blind or visually impaired. The participation of blind and visually impaired (BVI) individuals in STEM (Science, Technology, Engineering, and Math) fields has been notably limited (Beck-Winchatz & Riccobono, 2008). Among the many barriers that they face, existing computational thinking tools, particularly text-based programming environments, pose frequent accessibility challenges for BVI learners. For instance, screen readers may struggle to interpret code, and navigation via keyboard shortcuts is often inadequate. Similarly, block-based programming languages, such as Scratch or Blockly, rely on visual interfaces, making them largely inaccessible to BVI learners.

Traditional learning environments often rely heavily on visual information, but learning can be experienced through multiple senses, benefiting not only the BVI population but all learners. A hands-on approach, for instance, enables learners to create meaning and solve problems by interacting with tangible objects and technologies. In this project, we leverage hands-on learning in makerspaces—often referred to as '*maker learning*' or '*learning by making*', rooted in Seymour Papert's (1993) Constructionism—to engage BVI learners through its physical, tactile, and multimodal nature.

The goal of this project is to foster computational thinking (CT) skills in BVI youth, supporting their sense of agency, introducing STEM mindsets, and expanding career opportunities through maker learning. As Wing (2006) describes, CT is a universally applicable skill set and mindset for everyone, not just computer scientists, involving problem-solving, system design, and understanding human behavior by drawing on core computer science concepts (p. 33).

Ultimately, this project aims to create more inclusive and accessible learning environments for BVI learners and beyond. It is one of the few research projects conducted by, with, and for BVI learners, involving both blind and sighted researchers in *collaborative interdependence* (Bennett et al., 2018). The asset-based approach emphasizes the non-visual strategies and sensory abilities of blind individuals, such as touch, hearing, smell, and residual low vision, as central assets. Our blind researchers and participants play a pivotal role in leveraging these strengths to realize the full potential of multimodal learning, moving beyond vision-dominant tools and approaches to foster more accessible information, and learning environments for people of all abilities.

Research design

This research develops accessible maker tools and activities to foster computational thinking (CT) skills in BVI learners, while exploring the experiences of two key groups: (1) BVI learners and (2) librarians and maker professionals who design and deliver accessible computational thinking programs. This paper reports on the pilot phase, where the team delivered an accessible electronics and coding curriculum to three BVI youth in a two-day summer camp. Data was analysed from two debrief focus groups—one with BVI learners and one with maker professionals who served as instructors.

Tangible coding and electronics curriculum development

The accessible computational thinking program was co-designed by the research team, which included university researchers (one blind, one sighted) and graduate assistants, as well as three maker professionals (experienced coding instructors but new to teaching BVI learners). We created lesson plans around physical coding and electronics tools like Code Jumper, Snapino, and Snap Circuits, while adapting existing lessons to be more accessible for BVI youth. The activities focused on key concepts like circuits, coding loops, and conditional statements. The lessons

emphasized developing skills in tinkering, debugging, and collaboration, with a strong focus on tactile and auditory learning.

In hardware-focused activities, learners assembled circuits—including series and parallel—while exploring concepts like electrical flow, resistance, current, and voltage. Software activities using Code Jumper introduced programming basics by creating physical code sequences, encouraging both individual tinkering and group learning. The integration of hardware and software in tasks like using Arduino with Braille labels and Visual Studio Code allowed learners to apply coding concepts hands-on, using auditory and tactile feedback to enhance understanding, while visual elements were included but not essential.

Study setting, participants, and data collection and analysis

In August 2023, a two-day summer camp was held at the Chicago Lighthouse, an organization serving individuals who are blind, visually impaired, and disabled. The project team developed an accessible website and promotional materials in both braille and large print, with the Chicago Lighthouse supporting distribution and recruitment through their marketing tools. The University Institutional Review Board approved the project. Three blind or low-vision youth (2 males, 1 female, aged 15-20) participated in the camp, which included lessons on coding with Code Jumper, electronics with Snap Circuits, and hardware-software prototyping with Snapino. Instructors led the lessons, while researchers assisted as needed and conducted field observations.

After the camp, the research team conducted a focus group interview with each participant and instructor group. Each interview lasted approximately an hour and a half. All interviews were recorded and transcribed. The researchers conducted grounded theory approach to analyse the interview data.

Tables 1 and 2 below show information about the participating learners and instructors.

Learner Pseudonyms	Age	Race	Gender	BVI	Education
Imran	20	Asian/Pacific Islander	Male	Visually Impaired	College
Lilac	15	White/Caucasian	Female	Blind	High School
Isaac	15	White/Caucasian	Male	Blind	High School

Table 1. Demographic information of teen participants

Instructor Pseudonyms	Age	Race	Gender	CT Teaching Experience	BVI Teaching Experiences
Susan	36	White/Caucasian	Female	10 Years	None
David	30	White/Caucasian	Male	8 Years	None
Wyatt	28	Black/African American	Male	7 Years	None

Table 2. Demographic information of makerspace professionals

Findings and discussion

The findings present initial analyses from the pilot study, highlighting key insights into the experiences and perspectives of both BVI learners and instructors, along with instructional recommendations.

Effectiveness of the tactile and multi-sensory learning tools

Both learners and instructors noted that hands-on tools like Snap Circuit and Code Jumper enhanced learners' understanding of computational and electronic concepts. Overall, participants reported positive experiences with Snap Circuit and Code Jumper, while encountering challenges with Snapino. Snap Circuit helped make abstract concepts more tangible, and the use of a grid system with coordinate-based instructions further supported spatial understanding. Participants found this system particularly helpful for assembling circuits and understanding their functionality: "The braille and coordinates helped me understand how the fan worked. I've heard of using coordinates for chess but never with technology before."

> *Working with the circuits, I came to figure out the basics: positive, negative, current, the flow, and...resistance... Circuit was good, and we also did the music thing. I also learned that we can use the circuit and connect with the computer so we can manage the power, how much you want to send on the circuit... These were totally new to me. I didn't know any of them before. I did some coding in high school, but it was all on the computer, nothing physical. In this section, I guess it was practical. I could see what was happening physically. That was great. Things make more sense when you see them happening in physical form.*

Code Jumper helped learners grasp coding concepts by enabling them to '*feel*' how virtual codes are mapped onto physical block-based pods, each representing a specific function. This tactile interaction was useful for understanding the sequence and dependency relationships between different functions. Participants remarked:

each new peg—pause, loop, plus, minus—was interesting. The physical cords helped me understand how everything connected; *Having the physical cords and being able to tell how they're all connected, until there are too many pods to keep track of, really helped me understand. Each time we introduced a new element, like the pause or the loop and the different little pegs you plug in with* X *or plus/minus, those were really interesting.*

On the other hand, both learners and instructors reported challenges with the Snapino lesson taught on the second day, which was intended to build on the more engaging and playful learning experiences from the first day with Snap Circuits and Code Jumper. Snapino, a hardware kit that integrates Snap Circuits with the Arduino platform, allows users to work with both physical electronic components and programmable code (hardware-software integration).

However, teaching the program in the Arduino programming environment—a text-based coding platform—within a short period of time seemed to result in cognitive overload. Troubleshooting and debugging with the keyboard and screen reader were particularly difficult for participants. Debugging requires navigating the complex coding interface and using shortcuts, which were not fully covered in a brief, verbal-only session. While sighted participants can intuitively explore the user interface and interpret error messages visually, this same exploration posed significant challenges for BLV learners.

Learner confidence, autonomy, and interest

The self-reported learner focus group data highlight some of the key outcomes: increased confidence, autonomy, and interest. A participant who had previously had a negative experience

learning about circuits in a group setting at school mentioned that, with the use of braille and coordinate-based instructions, they were now able to understand how the circuit worked. The learner explained that their previous experience with computational thinking was not positive because their learning independence had been overshadowed by other sighted learners.

> *My teammates just did it for me because I didn't know what to do. Since it's a group activity, I couldn't really feel what was going on. I can't just touch it was going so fast. I could not just be like, 'Hold on, let me check it out by touching it.' They're just like, 'okay, we're going to do it.' Then it was just pretty confusing. Then there was a little quiz, and I was just like, 'I have no idea.' That was hard. But this time, I really liked that everything was described tactically, and nobody stepped in and took over or just did it for me…*
>
> *A lot of times I feel like people just take over, whether I'm doing a craft or whatever, they'll just be like, 'Oh, it looks like you're struggling over there.' Then they just do it. But then I wouldn't be able to put it together myself, so I really liked being able to put it together myself.*

Supporting learners' autonomy by providing tactile materials, hands-on guidance, and ample time for exploration appears to be an effective strategy.

Even though learners found the Snapino lesson quite challenging, a learner reported positive emotions, which fostered an interest in future learning. They expressed increased interest in coding after discovering the potential of using a keyboard and other assistive technologies for programming:

> *I feel like it's [Snapino] really cool that we got to use it on a real computer, and even though I wasn't very good at it, I think it just allowed me to see it is possible, which was really, really good. But I was a little bit too much of a beginner to actually problem-solve anything, so I just had somebody help me. But it was cool to see that you can use JAWS to read the code, and then use all those keystrokes to move around.*

Instructor challenges and strategies: verbal descriptions, adaptations, and personalized assistant

The participating instructors had extensive experience teaching STEM topics in informal learning environments, particularly in makerspaces, but they were new to teaching the BVI population. Prior to the camp, the project team, led by a blind researcher, provided a series of professional development sessions on accessibility. In multiple design sessions, including mock teaching role plays, the instructors wore blindfolds to better understand the challenges that BVI learners might encounter.

One of the major adaptations for instructors when transitioning from teaching sighted learners to BVI learners involved the extensive use of verbal descriptions. Instructors emphasized that verbal descriptions must be detailed, specific, and include tactile sensory information. They also suggested the use of analogies and metaphors to enhance understanding. These descriptions should precisely convey shapes, positions, and functions to assist BVI learners in identifying, navigating, and interacting with objects and their components. During the focus group, instructors noted the complexity and length of explaining Snapino to the BVI teens, which took approximately 10 minutes.

> *Instead of saying 'Grab **this** from your kit', you had to say 'Reach into the power bag, you're going to **feel around** for a **thicker** wire that's got a square end at one side and a **rectangular** end at one side. That rectangular **end is going to go into your computer**, make sure that this **part goes down** and the hollow end **goes up**, the square end is going to go into **the large metal** portion'.*

The instructors reported a learning curve in developing the skill to describe concepts and objects clearly and accurately without relying on visual references or awareness. They also noted the high cognitive demand placed on BVI learners, who had to remember complex instructions and spatial information. This required instructors to be especially mindful of their teaching pace and the complexity of their explanations, maintaining a balance between the number of descriptions given and the level of detail provided. Suggestions include using tactile teaching materials, such as tactile references to support self-guided navigation of the user interface and code. Supplementary learning materials could be created with a swell form machine to present the tangible layout of the user interface and code structure. Additionally, 3D shapes and references could be employed to further enhance understanding.

Finally, instructor focus group data emphasized the importance of understanding and accommodating the diverse needs of BVI learners. These learners have varying abilities, and what works for one may not work for another, requiring tailored approaches. Even within the small group of three learners, there were distinct needs. One learner had limited vision, braille literacy, and no additional disabilities. Another was completely blind and faced additional challenges, including mobility issues with assembling small components. The third learner had low vision, could read larger fonts, but had no braille literacy.

The instructors adapted their teaching in real time as challenges arose. For instance, during the Snap Circuit lesson, an instructor noticed that a learner with some vision, but no braille literacy was struggling to keep track of grid locations and made on-the-fly adjustments:

on the second day, I noticed he was struggling because [, without braille literacy,] he had to count and memorize grid locations. I gave him a new board with large letters, and it made a huge difference.

With a one-on-one instructor-learner ratio and real-time adjustments, the team was able to provide personalised assistance during the camp. However, challenges are anticipated in larger learning environments where such individualized support may not be as feasible.

Conclusion and future direction

This paper has demonstrated how tangible, accessible Maker tools like Code Jumper, Snap Circuits, and Snapino can be integrated into a coding and electronics curriculum through an initial analysis of the pilot study. The research team continues to analyse data collected during the camps, including pre- and post-survey responses, observation data, and triangulation with focus group data. The program tested in the camp has since been distributed to multiple public libraries and museum makerspaces. Two of these makerspaces engaged their local BVI learners, and we conducted interviews with the librarians who implemented the program. Ongoing data analysis aims to provide insights into multimodal learning and information processing, as well as efforts to sustain, improve, and disseminate the program more broadly.

While the pilot program shows promise and demonstrates the effectiveness of tangible coding platforms, a significant challenge remains in translating the skills acquired through these tools into text-based coding environments. Future work will explore how natural language processing and AI-powered programming assistants, such as GitHub Co-pilot, can help visually impaired individuals reduce unnecessary cognitive load. This '*speak-to-code*' approach may allow them to focus on core computational thinking processes rather than grappling with visual and textual interfaces. Additionally, emerging AI technologies, like 'Be My AI', are supporting BVI individuals

by describing visual items in everyday settings. Future research will investigate how such technologies can further reduce barriers for previously marginalized learners.

Acknowledgements

The project was made possible in part by the Institute of Museum and Library Services [LG-252360-OLS-22].

About the authors

Dr. Kyungwon Koh is an Associate Professor in the School of Information Sciences and the Director of the Champaign-Urbana Community Fab Lab at the University of Illinois at Urbana-Champaign, USA. She received her Ph.D. from Florida State University, and Her research expertise includes Library and Information Sciences, with a focus on youth services, learning, and community engagement; information behavior and literacies; and the maker movement in libraries. Contact email: kkoh@illinois.edu

Dr. JooYoung Seo is an Assistant Professor at the School of Information Sciences, University of Illinois Urbana-Champaign. He earned his Ph.D. from Penn State University, and his research focuses on accessible computing, inclusive STEM education, and multimodal data representation. Contact email: jseo1005@illinois.edu.

Dr. Si Chen is a Postdoctoral Researcher at the Lucy Family Institute for Data and Society and Notre Dame Learning at the University of Notre Dame, USA. She received her Ph.D. in information sciences from the University of Illinois Urbana-Champaign in 2024. Her research expertise lies in artificial intelligence in education and human-computer interaction, with a focus on how emerging technologies influence educational equity and accessibility through human-centered approaches. Contact email: schen34@nd.edu

Eugene Cox is a PhD student in Informatics at the University of Illinois at Urbana-Champaign. Their research interests include collaborative learning, learning sciences, and the use of technology to support collaborative learning. Their dissertation explores how secondary mathematics teachers use video clips to reflect on responsive teaching practices in middle school mathematics classrooms. Contact email: emcox3@illinois.edu

References

Beck-Winchatz, B., & Riccobono, M. A. (2008). Advancing participation of blind students in Science, Technology, Engineering, and Math. *Advances in Space Research*, 42(11), 1855–1858. https://doi.org/10.1016/j.asr.2007.05.080

Bennett, C. L., Brady, E., & Branham, S. M. (2018). Interdependence as a Frame for Assistive Technology Research and Design. *Proceedings of the 20th International* ACM SIGACCESS *Conference on Computers and Accessibility*, 161–173. https://doi.org/10.1145/3234695.3236348

Papert, S. (1993). Mindstorms: Children, computers, and powerful ideas (2nd edition). Basic Books.

Wing, J. M. (2006). Computational Thinking. Communications of the ACM, 49(3), 33–35.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Does Google dream of electric memes? From human to computational culture

Alexander O. Smith, Una Joh, and Jeff Hemsley

DOI: https://doi.org/10.47989/ir30iConf47350

Abstract

Introduction. We provide a cultural heritage informatic study of '*internet memes*': clusters of documents carrying a collective cultural meaning.

Method. Google's computer vision data from three memes are used to discuss the cultural heritage Google provides for the indexed web using Gini coefficients and word clouds.

Results. Our networks suggest that the more reducible sense-data is to a smaller number of labels, the less the internet needs to '*remember*' to see a '*meme*'. However, when labels are less strongly connected, human inspection is required to interpret if Google captures qualities of memes.

Conclusion. We show that Google can interpret aspects of memetic cultural heritage. However, it fails at understanding some cultural information about memes relative to predefined expectation. Computational tools like Google Cloud Vision can augment the heterogeneity of cultural heritage work. However closer inspection from experts is needed to interpret when computational methods represent cultural memory.

Introduction

Academic conceptions of '*internet meme*' typically assume a collection of files (e.g. jpeg or png files) that enact cultural memory: a defining interest of cultural heritage research (Assmann, 2008; Taylor, 1995; Viejo-Rose, 2015). As such, memes function as data of *cultural heritage informatics*: media imbued with the power of memory through iterative practices (Guan, 2008; Modrow & Youngman, 2023). Further still, memetics – the study of memes – studies dissemination, curation, classification, and organisation of culture (Dawkins, 1976; Shifman, 2014). In many cases cultural heritage informatics could easily borrow the name '*memetics*' without altering much. Yet, this sort of immediate association of research seems unlikely. Why?

The overlap seems conceptually intuitive, but underlying the intuition is a tension in disciplinary goals. Internet memetics regularly seeks to capture descriptions of '*new things*' in culture (Lankshear & Knobel, 2019; Smith et al., 2024): *disjunction*. Meanwhile cultural heritage seeks to find proper ways to canonise culture (Assmann, 2008) in relation to existing culture: *conjunction*. Internet memetics – despite conceptually being about memory and artefacts – is empirically resistant to the canonization of cultural connections by highlighting differences of interpretation (Lankshear & Knobel, 2019). Meanwhile, '*cultural heritage*' implies something inherited culturally. Cultural heritage pragmatically seeks to make '*something*' canon. Canonization of internet memes requires thoughtful responses within cultural heritage and memory institutions (Bratt & Smith, 2022; Rees, 2021; Tulloch, 2023), information studies broadly (Smith et al., 2024), but also within memetics (Pettis, 2021; Rogers & Giorgi, 2023).

In the internet research context, Pettis, Rogers, and Giorgi (2021; 2023) wonder how platform technology mediate, contextualise, and disjoin cultural memory. Instead, we turn to internet technology that conjoins the internet. Thus, we ask, '*How can Google search algorithms mediate, contextualise, and conjoin memetic memory*?' This question allows us to return to how to canonise the dynamics and heterogeneity of the collective memory of material culture. Thus, we redeem a fundamentally connected dimension of culture that memetic research seems to resist (Lankshear & Knobel, 2019) and the cultural dynamics of memory which *disciplinarily* defines heritage of cultural memory (Assmann, 2008).

To illustrate the issue, consider a brief motivating example. An '*internet meme*' is not a singular image-file nor a singular subjective memory. Rather, it is some arrangement (e.g., collection) of files, the way it moves through media, and the way it is remembered for every being who comes into contact with one or more of those image-files. Despite people frequently having claimed to have '*seen/made/shared a meme*,' this reduces a meme to a subjective interpretation of singular, isolated files. Likewise, saying '*I have curated a meme*' renders the meme dynamics invisible in the way they are technically and cognitively existing. Either case renders a meme without the context of how it is a meme, failing the conceptual objective of memes.

To properly '*curate a meme*' requires representing a meme inclusive of dynamic cultural meanings, dimensions, and implications. Researchers and cultural heritage practitioners must inseparably and accurately represent both (1) dynamic cultural memory and (2) an accurate description of its digitally mediated context.

This paper approaches the dynamics of memory for static image-files and the technologies that interpret '*memory*' for us as a kind of cultural heritage. Especially for memetics online, understanding how internet technologies interpret memetic content will shape the way we humans interpret culture within iterative expressions that emanate from memory practices. One of the most powerful technologies for the human interpretable internet is web search. Google, in particular, remains the most recognized search engine online (Bianchi, 2024). Understanding how Google *conjoins* the image-files that make '*a meme*' empirically closes a gap in connecting memes' static files to cultural heritage. In particular, the data that establishes regularities in web search

and information retrieval are systematically organized through iterative, aggregative, networked, and collective habits of search. While perhaps one might suggest that '*humans*' are absent in studying Google search, it is data humans 'invisibly' produce through search that provides a significant portion of Google's regularized data for information retrieval.

Literature review

Memetics is usually considered a kind of datafication of cultural information (Smith et al., 2024). Richard Dawkins (1976) coined the term meme to describe a unitary bit of cultural information which inherited from '*mind-to-mind.*' He exemplifies this through architectural '*arches*' as a cultural artefact in the world which is fundamentally connected to the mental representation of an '*arch*'. Later, Daniel Dennett integrated this into his theory of mind (1991/2017). Dennett illustrated how cognitive connections pass these idealised bits of culture by way of information, using a myth about how Marco Polo brought pasta from China to Italy through memory and documentation.

These early accounts of '*meme*' were critiqued and re-theorised for digital contexts by Limor Shifman (2014) and others (see, Smith et al., 2024). Yet, memes do actively entangle artefacts and memory (Smith & Loewen-Colón, 2024), but few researchers have empirically shown how '*internet memes*' and memory connect.

Often '*mental*' perspectives – cognition, phenomenology, memory, knowledge, and so forth – are buried just under the data of technical objects (Rogers & Giorgi, 2023) or communication data (Shifman, 2014). Yet, meme's information conjoins minds, albeit with more nuanced and multi-mediated ways than Dawkins' or Dennett's reductive '*mind-to-mind*' descriptions. As contemporary memetic research clarifies, there are often a plurality of information technologies between '*mind-to-mind*' mappings.

To redeem mental processes of memetics, how can we begin to understand what '*memory*' is when the data are outside of human minds, and instead are in the world as artefacts? Previous research has suggested a relationship between cultural heritage and memes (Demsky, 2021; Modrow & Youngman, 2023). However, empirical interests of memetic cultural heritage (Demsky, 2021; Petrassi, 2024; Rees, 2021; Tulloch, 2023) usually investigate human-centric experiences without the nuance of technical mediations that conjoin minds. As such, there is a gap between internet memetic technological arguments and cultural heritage's research on individual human experiences. Complex digitally mediated ecologies of memes (Shifman, 2014) and cultural heritage need to be re-joined to explain memetic memory (Smith & Loewen-Colón, 2024).

When memetic data are collections of image-files, the cultural heritage task is clarified. Relatedly, memory activates through these files. Despite being corollary, digital files and mental processes are not reducible to each other. Shifman (2014) suggests '*social mindsets*' entangle inseparably with these files but remains open to how this occurs. She defines memes as '*content, form, and/or stance*' which must travel across the internet by way of many users. Content is the '*idea*,' form is the meme's '*media*', and stance is how the information enables unique contextual meanings for individuals as the meme is communicated. Despite the intentional entanglement of these concepts, internet memetics have tended towards turning '*content*' of a meme into a plurality of meanings (Lankshear & Knobel, 2019; Smith et al., 2024) which tempts researchers into reducing Shifman's definition to merely digital media's immediate context (Wiggins, 2019). Doing so renders the transmission and dynamics of cultural heritage characterised within Shifman's definition inaccessible. To redeem memory, Smith, and Loewen-Colón (2024) create initial steps in re-theorising memory for both Shifman and Dawkins.

While lots of theorisation of transmission of memetic information exists (Cappella et al., 2015; Dawkins, 1976; Hull, 2000; Smith & Hemsley, 2022; Smith & Loewen-Colón, 2024), observing

transmission in digital contexts is a difficult task without near exact copies of artefacts such as '*retweets*' or '*share*' buttons (Sperber, 2000). This is partly why '*viral events*' were distinguished from '*memes*' in digital contexts (Nahon & Hemsley, 2013; Shifman, 2014). However, information transmission and meme might have nuanced connections if we could conceptualise what information constitutes a meme in some statistically regular image-file data. To fill the empirical gap between materiality of image-files and the '*content*' of memetic memory, we must start by seeing information selection across image-files.

Theoretically, the relation of these image-files and memory have been recently updated as being connected to '*traces*' (Smith & Loewen-Colón, 2024). Information selection of digital and cybernetic technologies are often called *trace data* (Geiger & Ribes, 2011; Østerlund et al., 2020). Trace data signifies movement or *selection* (Cappella et al., 2015; Heylighen, 1993) of information. Studying memetic traces empirically remains incompletely described, yet previous literature has plenty of suggestions.

Prior work suggests interpreting memetic data traces are digitally mediated and technologically situated (Rogers & Giorgi, 2023; Smith et al., 2022). However, Smith et al. suggests trace data can be understood through the datafication of Google image web search technologies (Smith et al., 2022; Smith & Hemsley, 2022) and across digital technologies more generally (Smith & Loewen-Colón, 2024). Google both provides traces of web indexing, but also artefactual traces '*within*' each artefact. We distinguish *artefact* from jpeg or png files: memetic *artefacts* are data commonly identified across files. When Google identifies a frog or a cat in two different image-files, that frog or cat operates as *artefactual trace data* which Google considers culturally relevant for web search, indexing, and so forth. Google '*traces*' a cat to a bunch of other cat artefacts. If Google considers this data relevant for web search, then it is this data that informs our cultural experience online.

By using Google as a method of recovering dynamics across image-files by tracing artefacts across them, we ask the following questions:

> RQ1. What do Google's digital traces suggest about the conjoining of memetic image-files as technologically mediated cultural heritage? How are different meme's image-files similar and/or different?
>
> RQ2. Do Google's digital artefact relations seem correlated to the digital traces provided by a memetic collection of image-files on their own through our initial interpretation?

RQ1 considers technologically mediated conjunction using computational network methods and compares across three qualitatively selected memes based on what the authors deemed structurally different kinds of traces. RQ2 considers these three structurally different meme image-file collections to interpret similarity and differences (Smith & Hemsley, 2022) of artefactual traces across the files. The first question considers what the technology conjoins as the regular semiotic '*structure*' of each collection. The second asks what we expected would be conjoined in relation to what actually was conjoined. Together these questions suggest how technologically mediated conjunction might be different from expected conjunction. We will explain how we provided these differences across three memes in the next section.

Methodology

With interest in what data traces that sustain collective/distributed internet memory, we use relevant technologies that provide this datafication following Smith et al. (2022) who use a cyber-semiotic interpretation (Cannizzaro, 2016) of memetic traces. As such, our research design makes use of Google data gathered from Google Cloud Vision's API. This provides us data that informs Google about what image files are affiliated with text search and reverse image search. These data play a large role in popular web search, shaping our digital cultural environment.

However, we are also interested in how a '*meme*' that people curate into a collection of image documents '*belong together*'. For this we leverage the cultural '*authority*' of Know Your Meme's (KYM's) image collections as other researchers have done (Katz & Shifman, 2017; Smith et al., 2022; Smith & Hemsley, 2022). KYM curates image-files into thousands of '*meme entries*,' each of which operate as cultural heritage collections for '*a meme*.' The following section shows how we use KYM's meme entries to gather cultural heritage data from Google.

Data collection

Before getting data from Google, we downloaded three meme entries: Doge, Hello There, and Loss. Doge typically contains a Shiba Inu dog in a variety of contexts. '*Hello There*' is a meme based on character dialogue in Star Wars between Obi-wan Kenobi saying, '*Hello There*' and General Grievous responding, '*General Kenobi*!' However, as can be seen below they are somewhat visually correlated, but they are not as ubiquitous in their visual data as Doge. Loss is a pattern '| || || |_' in some literal or translated way. Example image-files of these meme entries can be seen in respective order in Figure 1 below:

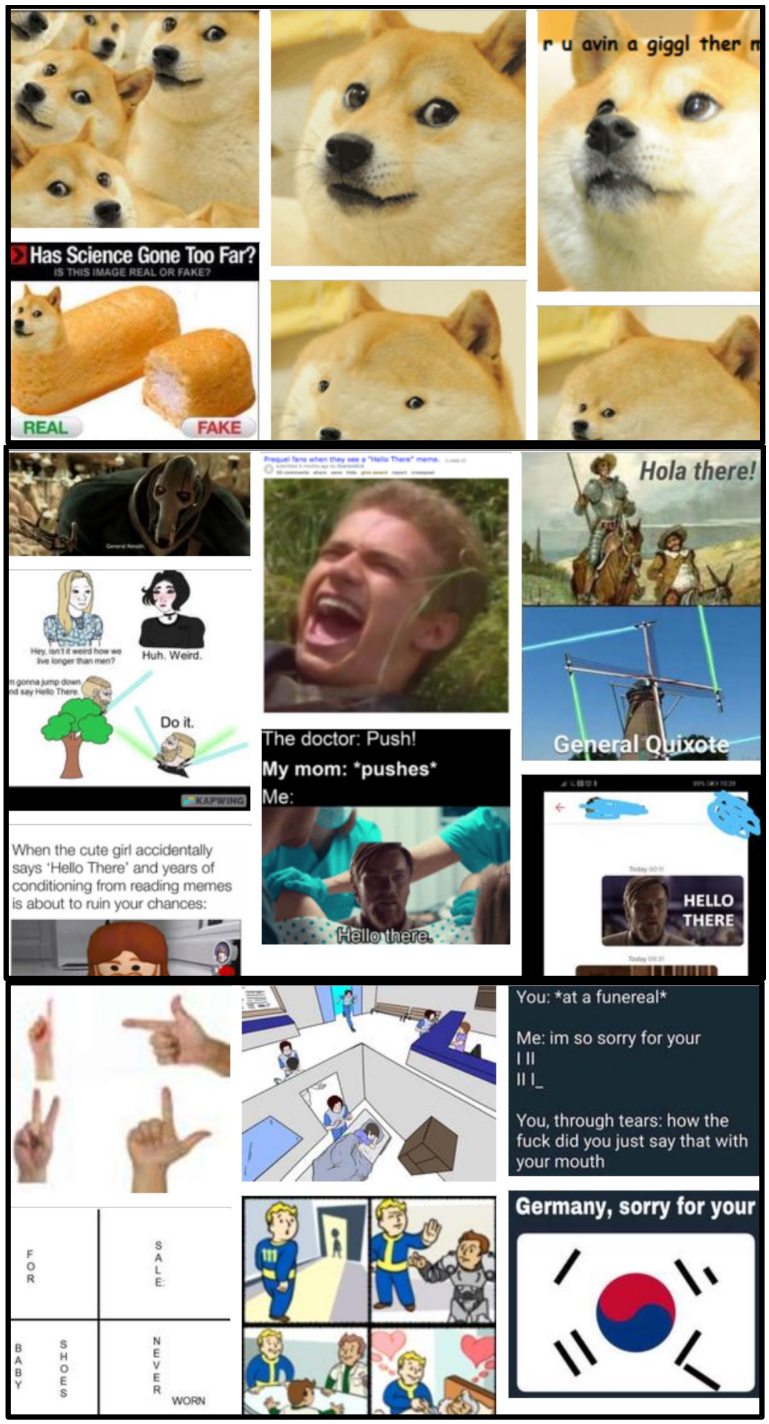


Figure 1. Examples of Doge, Hello There, and loss screen captured from KYM

These were chosen because of visual differences and the way they are culturally remembered through traces that conjoin the image-files together. Through visual inspection, the authors noted the collection of image-files appear to belong together. '*Doge*' is a collection of files which almost exclusively belong to the collection because they have the iconic face of a photoshopped Shiba Inu dog that is mostly used as a reaction image. It is often used to signify an emotion like an emoticon. '*Hello There*' image-files each contain some subset of a larger set of regular co-occurring artefacts within each file. Finally, Loss is unique in that it contains a geometrically abstract pattern. While all the images contain this pattern, the pattern itself seems to require abstract imagination to

observe it. While it might be obvious that Google can observe a Shiba Inu and assume an association to '*Doge*' in text search, assuming an abstract pattern like Loss would be akin to searching for '*Hello There*' in Google and expecting that the text alone will return lots of Star Wars results.

While there are perhaps many '*different kinds*' of memes than this, we could not find a complete taxonomy of memes that tries to interpret them through inspections of trace data, although it seems methodologies which might enable this exist (Smith & Hemsley, 2022). Although there are highly context dependent taxonomies of things like '*kinds of memetic nonsense*' (Katz & Shifman, 2017) which are not within the scope of the present study's goal.

Thus, after having collected the three meme entries, we used Google Cloud Vision API to return the relevant image-file features. Our relevant *features* were '*labels*' and '*web entities*'. *Labels* are tags that Google's cloud vision attaches to images. For example, if Google sees a Doge affiliated image-file, it will likely return the label '*dog*'. Labels are more immediately observed visual artefacts within the image file and serve more as immediate sense data. *Web entities* are search terms or Google's '*inferences*' that are mostly likely associated with the images in web search. For example, if Google recognizes a dog in a Doge image, then likely it will return web entities such as 'Shiba Inu', '*Doge*', etc that are more indicative to cultural heritage, and thus the connective '*memory*' of the image file. Labels proxy as '*sense-data*' while web entities proxy for technologically sustained '*memory*': i.e., the cultural heritage sustained by Google.

Data analysis protocol

After having collected the image-files from the three meme entries and the Google Cloud Vision data, we have a set of human readable data that reflects what visual artefacts are in each image document, but also aggregately within each meme collection. To interpret what artefacts, get '*selected*' within each meme collection, we interpret the labels and the web entities as '*nodes*'. However, instead of including all the detected features as nodes, we choose features that have more than a 0.5 confidence score, which shows how certain Google Cloud Vision is about the detection of that feature. From these nodes, we construct two sets of networks of co-occurring features: one for labels, and another for web entities for each meme entry.

In feature-to-feature networks, the weight of an edge represents the number of common images-files where two labels or web entities appeared together. This is useful because, low co-occurrences, such as those that only occur once (edge weight = 1) can be weighted as less important to a meme entry than those that occur within 10 images (edge weight = 10) for example. Weighted graphs enable us to calculate the '*strength*' of nodes, a metric that is the sum of the weights of the edges connected to a node (Barrat et al., 2004). This serves as an alternative measure of centrality of feature data to a meme entry, accounting not only for the number of connections, as in degree centrality, but also for the intensity or importance of those connections. The strength of the features enables us to '*quantify*' the visual similarity and differences between two images and help us describe the 'fuzziness' of semiotic references as we understand them in internet memes.

From the strength value of each node, we calculate the Gini coefficient. The Gini coefficient is a measure of inequality in a distribution, ranging between 0 and 1, where 0 corresponds to perfect equality. In our networks, the Gini coefficient would be close to 0 if all the features have the same strength, and close to 1 when there is a large disparity among the strengths. Even though each meme entry has a different number of nodes (the number of features that appear in the images of each meme entry), we can still compare the Gini coefficients because it is a relative measure of inequality that is independent of the network's size.

Results

There are two types of networks: (1) feature networks and (2) image networks. Feature networks represent how the features (either label or web entities) are connected to each other based on common images that include both entities. Conversely, image networks show how individual images-files are connected based on common entities that are detected in both images.

Feature networks

We calculated the strength of all the nodes in the network, which was used to produce a Gini coefficient by using a '*Lorenz curve*' (Hu & Wang, 2005; Kirillov & Panov, 2022). The Lorenz curve provides the basis for calculating the Gini coefficient by using the ratio of the area between the Lorenz curve and the line of perfect equality to the total area under the line of perfect equality. The resulting Gini coefficient allows us to compare inequality in the distribution of strength values. That is, a lower Gini coefficient suggests that strength values are quite evenly distributed across features. Conversely, a higher Gini coefficient shows that a few nodes (features) have very high strength values—meaning they share a lot of commonalities with others—while most nodes have lower strength values. Then the Lorenz Curve was produced by ordering the ratio of nodes by their strengths, seen in Figure 2 for each meme.

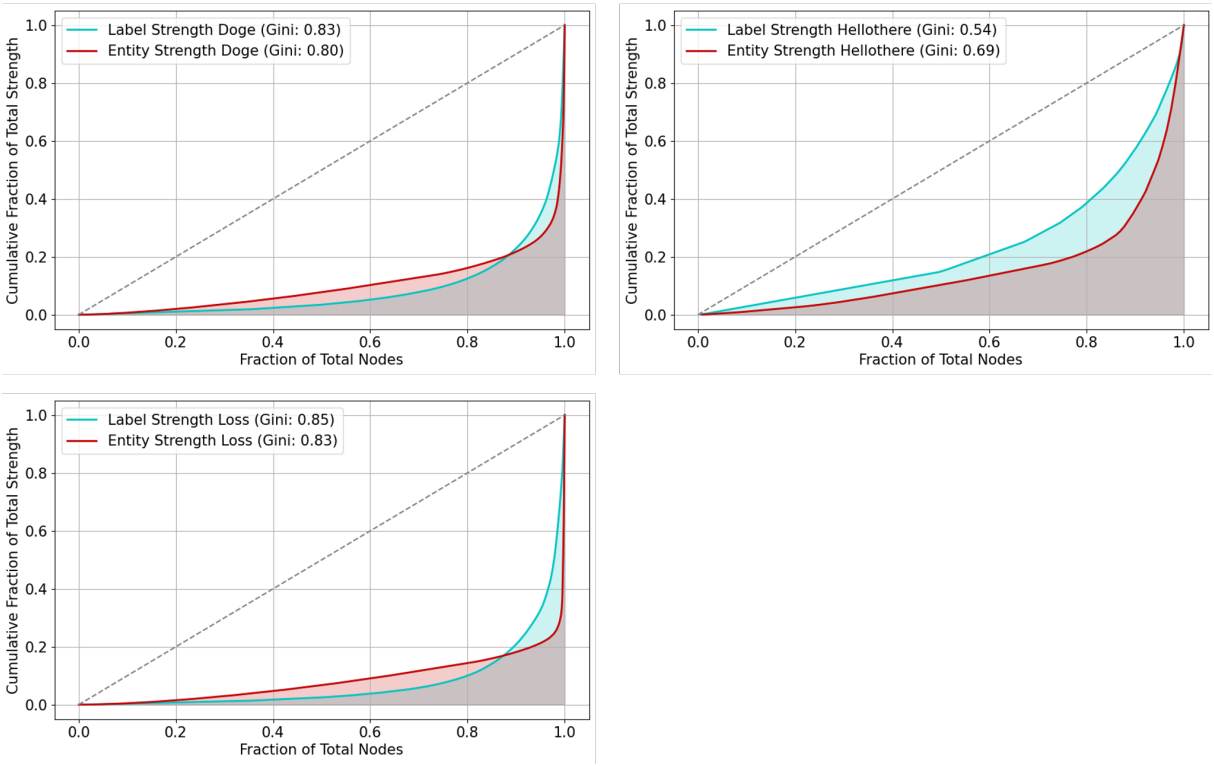


Figure 2. Lorenz curve and Gini coefficient of node strength: comparison between labels and entities

What Figure 2 shows us is a comparison of each meme's label and entity features' Gini coefficient. As the figure shows, Doge and Loss are relatively more hierarchical in their labels than their entities. That is to say when Google interprets the labels, it extends those labels to entity data that are less hierarchical. This means Google '*sees*' the image-files as having more specific visual features, but it has less '*memory*' of what those image-files are about specifically. The opposite is true of Hello There. Google 'sees' the image-files of Hello There, and 'remembers' more specifically what it is about.

Answering RQ1, we find that Google captures the cultural heritage of Hello There in higher specificity than Loss or Doge because of how it extends to more specific cultural contexts with the

strength of particular entity labels' meanings to the meme. This can be seen in a word cloud for Hello There. We see that from the more dominant labels, there are a lot of references which may or may not obviously relate to the cultural '*meaning*' of Hello There as a meme. However, as we move to the entities, the cultural context of Hello There becomes more clear in that Google sees that it is related to Star Wars and its characters for example.

However, Doge and Loss, from Figure 2 appear remarkably similar. This is counterintuitive based on our visual inspection. In order to distinguish if Google actually sees and '*remembers*' them differently we look at what it sees relative to what Google remembers by considering the entities names themselves. We can also extend what we just saw from Hello There and see what happens to the most '*strong*' features as we move from Doge's and Loss' labels to entities in Figure 3, below.

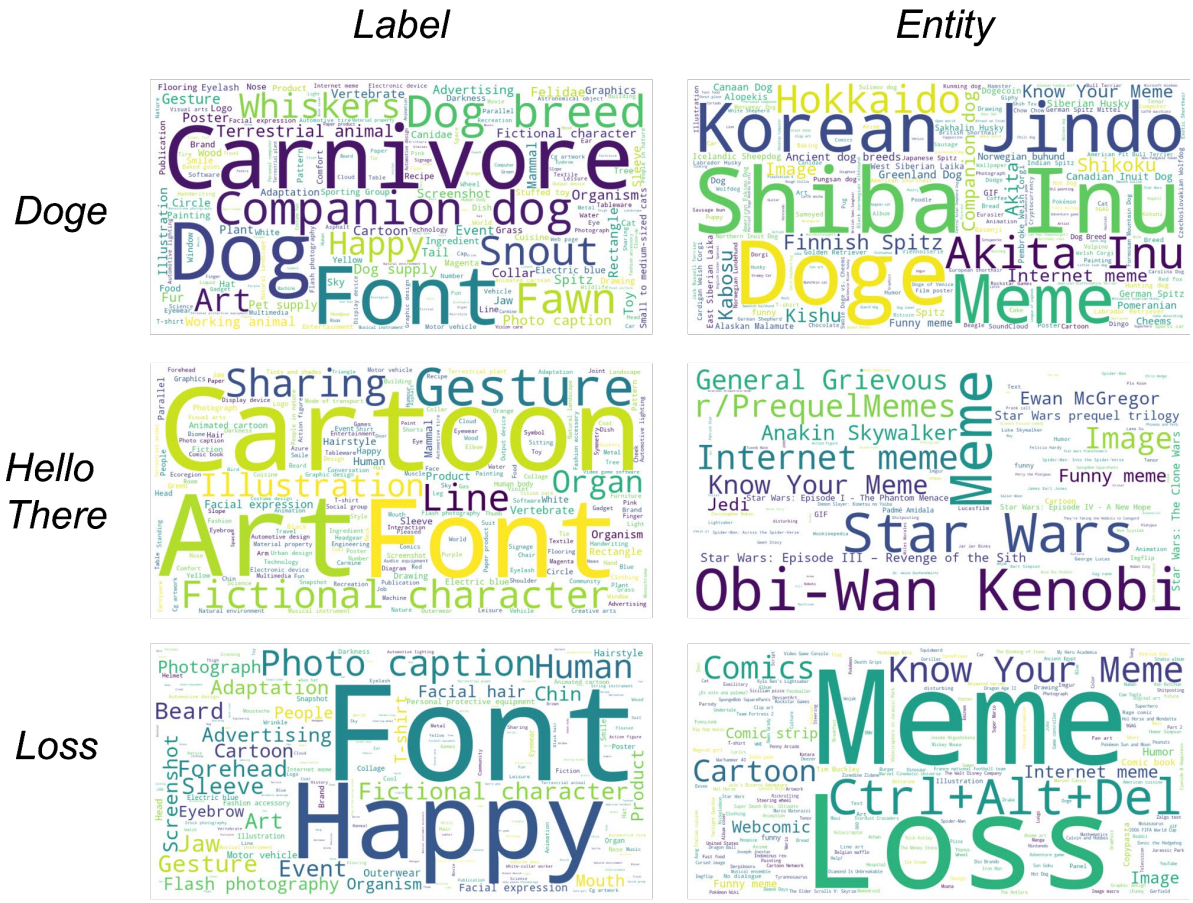


Figure 3. Word cloud by strength of feature

This figure suggests that for Doge, the entities it remembers are much closer to what we expected. Google '*sees*' text and a dog, and '*remembers*' it is a 'meme' of a '*shiba inu*' called '*Doge*'. That is, while perhaps Google knows more of what '*Doge*' is informationally about based on cultural expectations. It remembers that it is seeing a Doge meme and many different aspects core to Doge, which is the core of the meme. At first glance, it appears as though the same is true for Loss. However, recall what makes Loss culturally relevant is more about a pattern than what we see here. While Google '*remembers*' it is a meme called '*Loss*' it captures more information about its source material: Ctrl+Alt+Del, a cartoon or webcomic. Or alternatively it particularly generically refers to Know Your Meme in a nonspecific way. If it captured what makes Loss, the meme entry, '*belong together*,' one of the primary words would be something like '*pattern*' or some text representation of Loss: '| || || |_'. Google trace data's lack of mapping to human memory of what makes Loss culturally

specific is what was expected. It is this way that Doge and Loss are different from each other in Google's sustained trace data of these memes. Google's data suggests it knows what is culturally relevant about Hello There and Doge in Figure 3. However, Figure 3 shows us that the critical bit of information that makes Loss different in the kind of meme it is from the other two is not captured. Google merely reiterates generic data about the meme. This answers RQ2.

Discussion

Hello There

Our results do suggest strong reasons to think that Hello There is significantly different from Doge or Loss. While Doge and Loss are consistently extremely hierarchical, Hello There is consistently the opposite. In terms of Google's features, this is because of the combinatorial arrangements of what allows something to fit into '*Hello There.*' As such, there is more cultural content to describe Hello There. As we see in the Figure 3, despite there being more cultural content, it seems to summarise what we expected. Hello There comes from a narrative interpretation provided from Star Wars that has regularly been translated into other languages, narrative contexts, and iconographic aspects of Star Wars. That is, Google 'sees' things that are highly correlated to Star Wars, and thus Google connects the correlates as a regularised '*digital memory*'. In our authorial minds, it is intuitive that the meme entry would have observable connections to the narrative. As our Gini coefficient suggests, it is that when the labels are organised together, somehow Google algorithmically predicts relations to web entities based on digital regularities. Google expects those labels to extend towards terminology that is associated with Star Wars, lightsabers, Obi-Wan Kenobi, General Grievous, and so forth. So, the web entities focus on aspects of the Star Wars story in its slightly more hierarchical interpretation of the meme. Since we '*know*' the story of Star Wars, this is sensibly intuitive because the data we typically remember are digitally representable. However, the actual computational method of doing this is inaccessible to us, but it must be computational and algorithmic, and thus digital. So, while it requires a large number of references for a digital cultural actor to understand Hello There as a human cultural phenomenon. However, the word cloud shows that it does not remember the '*narrative*' call-response, '*Hello There*', followed by '*General Kenobi!*' which is what the meme ultimately is about. So, while it focuses on the surrounding narrative, it does not seem to recognize it fully.

Doge

Doge's word cloud also maps well to our memory of it based on our interpretation of what is core to Doge: The Shiba Inu and surrounding characteristics. As a meme that's about bodily gesturing and facial characteristics, it captures something very similar to that. However, we noticed the Gini coefficient suggests Google '*remembers less*' about the meme. This is perhaps because Doge has '*less*' information about it other than the particular dog in the image. Doge culturally floats into and out of many different cultural contexts. People post it everywhere with no particular '*linguistic*' goal, but rather it has an expressive goal like how someone might use an emoticon. This is contrary to '*Hello There*' which is a call-response sort of joke from Star Wars. That is, we know Google knows Doge because the Doge meme culturally is a dog and not much else: a floating signifier. '*Hello There*' is known through the nodes that strongly hold it together: the combinatorial features. However, the word cloud suggests that it does not understand the '*call-response*' cultural aspect of the meme.

Loss

Loss, as expected, is particularly difficult to capture by Google both by Gini coefficient and by what we culturally expected to be in the word cloud. This is because what makes 'Loss' work is abstract, and difficult to reduce to language. Simply networking the image-files together creates a less strong, cohesive idea of what the meme centres around. And, as such, it remembers the source

material and where the meme is mostly documented more so than what the meme means or is about. Google merely repeats its context back to us without content.

Conclusion

In total, we see that Google's '*memory*' of memes as collections of image-files in some cases performs better at capturing what we expected of memetic cultural heritage, and sometimes it does not. In particular, Google understands what '*Hello There*' and '*Doge*' are about, however we have to see these two memes in different ways. Hello There, is understood more as a network of features by Google where Doge only seems to be understood through the dog, but what holds the meme together otherwise is not known to Google. Loss, however, seems to lose context in both cases. Essentially Google thinks '*Loss*' is a webcomic or a generic meme on Know Your Meme called Loss. It does not give us any indication of its cultural relevance.

What our study suggests is that certain parts of cultural memory are not reducible to computationally reducible visual data. More context is needed to understand what makes a meme important. We see this in each of our meme's network data. However, we do see that Google can provide a partial understanding of what artefactual trace data frequently ties them together. As such, we can begin to see what information selection is occurring across collections. However, in the case of Hello There and Doge, it seems Google cannot point to exactly what gets selected. That is, Google's attempt to describe Doge leads it to various descriptions of Doge and not the cultural contexts in which Doge is used. It is even less precise about Loss. On the other hand, Hello There is more precise in its network, but it is less precise about its narrative content other than it is '*funny*'.

In relation to memetics, we have seen that some memes are understood through computational artefact networks through Google vision features. However, we have also given some insights into where it is limited. For example, it does not understand when cultural context is not within the data, or when the cultural context is strongly implied through probabilistic references in the case of Loss and Hello There.

Our literature review suggests that Google can actually play a significant role in what gets sustained in cultural heritage, and thus can shape cultural memory. As such a significant part of our cultural knowledge, how it understands or misunderstands this culture plays a part in a computationally sustained cultural heritage as trace data. Memes, when understood as '*particulates*' of documented culture, provide a way to be granular about understanding computationally driven tools like Google Cloud Vision's usefulness or limitations in sustaining the relevant parts of cultural memory. In order to be clear about what Google generates for cultural heritage, connections between cultural expectations of those that '*know*' and an understanding of how the qualitative aspects of cultural heritage are bridging with computational tactics of making sense of them. While our results suggest that Google is '*good*' at understanding certain parts of memetic meaning, we find that it needs to be augmented through human interpretation to make sense of a lot of what it provides. Google, in particular, can significantly aid in bridging new cultural heritage together however, since it pulls together a lot of information in the heterogeneity of the indexed web, so understanding such tools helps augment limited subjective views of cultural heritage work. Future work is needed to understand what computational aspects can be fully leveraged in augmenting cultural heritage work and what should be delegated to humans.

About the authors

Alexander O. Smith is a PhD candidate of Information Science and Technology at Syracuse University. Their research is focused on memetics: the study of memes. Alexander is forwarding a novel theoretical framework for memetics which synthesizes older memetic theories with internet

meme studies. Affiliated work can be found at *Convergence*, *Journal of Documentation*, ASIS&T, and the IEEE big data conference computational archive science workshop. More expansive research interests on (cross-)platform dynamics, network cultures, and classification can be found at *Policy and Internet*, *Social Media + Society*, eCSCW, and AoIR.

Una Joh is a Ph.D. student at the School of Information Studies at Syracuse University, specializing in computational social science. With a background in economics from her bachelor's and master's studies at SKKU in Korea, she conducts research on the flow of information and the dynamics of public opinion on social media, taking into account its interaction with news media. Her research aims to extract, track, and predict information flows and opinion dynamics on networks that model interconnected media environments. To achieve this, she utilizes a variety of methods, including natural language processing, network analysis, and agent-based modeling.

Jeff Hemsley is the Interim Dean at the School of Information Studies at Syracuse University. He earned his Ph.D. from the University of Washington's information school. His research is about understanding information diffusion, particularly in the context of politics, in social media. He is co-author of the book Going Viral (Polity Press, 2013 and winner of ASIS&T Best Science Books of 2014 Information award and selected by Choice magazine as an Outstanding Academic Title for 2014), which explains what virality is, how it works technologically and socially, and draws out the implications of this process for social change.

References

Assmann, A. (2008). Canon and Archive. In A. Erll & A. Nünning (Eds.), Cultural Memory Studies: An International and Interdisciplinary Handbook (pp. 97–108). De Gruyter. https://doi.org/10.1515/9783110207262.2.97

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 101(11), 3747–3752. https://doi.org/10.1073/pnas.0400087101

Bianchi, T. (2024). Online search market worldwide—Statistics & Facts (Online Search). Statista. https://www.statista.com/topics/1710/search-engine-usage/

Bratt, S., & Smith, A. O. (2022). Evolutionary Archives: The Unlikely Comparison of GenBank and Know Your Meme. 2022 IEEE International Conference on Big Data (Big Data), 2490–2499. https://doi.org/10.1109/BigData55660.2022.10020929

Cannizzaro, S. (2016). Internet memes as internet signs: A semiotic view of digital culture. Σημε$ιota$ømega$τκ$\acute\eta$-Sign Systems Studies, 44(4), 562–586. https://doi.org/10.12697/SSS.2016.44.4.05

Cappella, J. N., Kim, H. S., & Albarracin, D. (2015). Selection and Transmission Processes for Information in the Emerging Media Environment: Psychological Motives and Message Characteristics. Media Psychology, 18(3), 396–424. https://doi.org/10/gf8x87

Dawkins, R. (1976). The Selfish Gene. Oxford University Press.

Demsky, J. (2021). That Is Really Meme: Nazi Pepe the Frog and the Subversion of Anglo-American Holocaust Remembrance. In J. Demsky (Ed.), Nazi and Holocaust Representations in Anglo-American Popular Culture, 1945–2020: Irreverent Remembrance (pp. 105–125). Springer International Publishing. https://doi.org/10.1007/978-3-030-79221-3_7

Dennett, D. C. (2017). Consciousness Explained. Little, Brown. (Original work published 1991)

Geiger, R. S., & Ribes, D. (2011). Trace Ethnography: Following Coordination through Documentary Practices. 1–10. https://doi.org/10/fqntnc

Guan, K. C. (2008). Oral Histories in the Making of Our Memories and Heritage. https://doi.org/10.1163/156853108X327100

Heylighen, F. (1993). Selection Criteria for the Evolution of Knowledge. Proc. 13th Int. Congress on Cybernetics (Association Internat. de Cybernétique, 524–528.

Hu, H.-B., & Wang, L. (2005). The gini coefficient's application to general complex networks. Advances in Complex Systems, 08(01), 159–167. https://doi.org/10.1142/S0219525905000385

Hull, D. L. (2000). Taking memetics seriously: Memetics will Be what we make it. In R. Aunger (Ed.), Darwinizing culture: The status of memetics as a science (pp. 43–67). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780192632449.003.0003

Katz, Y., & Shifman, L. (2017). Making sense? The structure and meanings of digital memetic nonsense. Information Communication and Society, 20(6), 825–842. Scopus. https://doi.org/10.1080/1369118X.2017.1291702

Kirillov, B., & Panov, M. (2022). Measuring internal inequality in capsule networks for supervised anomaly detection. Scientific Reports, 12(1), 13575. https://doi.org/10.1038/s41598-022-17734-7

Lankshear, C., & Knobel, M. (2019). Memes, Macros, Meaning, and Menace: Some Trends in Internet Memes. Communication and Media Studies.

Modrow, S., & Youngman, T. (2023). Theorizing Cultural Heritage Informatics as the Intersection of Heritage, Memory, and Information. Proceedings of the Association for Information Science and Technology, 60(1), 666–671. https://doi.org/10.1002/pra2.836

Nahon, K., & Hemsley, J. (2013). Going Viral (1 edition). Polity.

Østerlund, C., Crowston, K., & Jackson, C. (2020). Building an Apparatus: Refractive, Reflective, and Diffractive Readings of Trace Data. Journal of the Association for Information Systems, 21(1), 1–22. https://doi.org/10.17705/1jais.00590

Petrassi, D. (2024). Memetic Aesthetics between Artistic Production and Cultural Heritage. This is why we must consider memes the multimodal artifacts of the 21st century.. Southern Semiotic Review Issue, 19. https://doi.org/10.33234/SSR.19.3

Pettis, B. T. (2021). Know Your Meme and the homogenization of web history. Internet Histories, 0(0), 1–17. https://doi.org/10/gmkh47

Rees, A. J. (2021). Collecting Online Memetic Cultures: How tho. Museum and Society, 19(2), 199–219. https://doi.org/10/gmfdcw

Rogers, R., & Giorgi, G. (2023). What is a meme, technically speaking? Information, Communication & Society, 0(0), 1–19. https://doi.org/10.1080/1369118X.2023.2174790

Shifman, L. (2014). Memes in Digital Culture. The MIT Press.

Smith, A. O., & Hemsley, J. (2022). Memetics as Informational Difference: Offering an Information-Centric Conception of Memes. Journal of Documentation, 78(5), 1149–1163. https://doi.org/10.1108/JD-07-2021-0140

Smith, A. O., Hemsley, J., & Tacheva, Z. Y. (2024). Reinforming memes: A literature review of the status of memetic information. Journal of Documentation, 80(4), 1003–1021. https://doi.org/10.1108/JD-10-2023-0218

Smith, A. O., & Loewen-Colón, J. (2024). Memetic memory as vital conduits of troublemakers in digital culture. Convergence. https://doi.org/10.1177/13548565241262421

Smith, A. O., Tacheva, J., & Hemsley, J. (2022). Visual Semantics of Memes: (Re)Interpreting Memetic Content and Form for Information Studies. Proceedings of the Association for Information Science and Technology, 59, 800–802. https://doi.org/10.1002/pra2.731

Sperber, D. (2000). An Objection to the Memetic Approach to Culture. In R. Aunger (Ed.), Darwinizing Culture: The Status of Memetics as a Science (1st ed., pp. 163–173). Oxford University Press.

Taylor, H. A. (1995). "Heritage" Revisited: Documents as Artifacts in the Context of Museums and Material Culture. Archivaria, 8–19.

Tulloch, B. J. (2023). Memes to an End: Why Internet Memes Matter to Information Research. Libri, 73(4), 293–305. https://doi.org/10.1515/libri-2021-0098

Viejo-Rose, D. (2015). Cultural heritage and memory: Untangling the ties that bind. Culture & History Digital Journal, 4(2), e018. https://doi.org/10.3989/chdj.2015.018

Wiggins, B. E. (2019). The discursive power of memes in digital culture: Ideology, semiotics, and intertextuality. Routledge.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Teaching computational archival science: context, pedagogy, and future directions

Victoria L. Lemieux, and Richard Marciano

DOI: https://doi.org/10.47989/ir30iConf47347

Abstract

Introduction. The paper describes the development of the new transdisciplinary field of computational archival science (CAS) and the integration of computational thinking (CT) concepts into archival science.

Method. The authors show how CAS can be introduced into graduate archival training through two case studies at the University of Maryland and the University of British Columbia and discuss building and sustaining CAS educator networks.

Analysis. The paper argues that, given the increasing use of AI in archival work and research, the acquisition of computational skills and competencies is urgent for those entering the profession, but sees several barriers, including the willingness of archival educators to engage in this space, and the shortage of CAS educators. There is also a perceived conflict among some in the archival profession between CAS and recent archival scholarship emphasizing postcolonialism themes.

Results. Results show this is a false dichotomy, as demonstrated by the many CAS papers focusing on ethical and social justice aspects of computing and archival work.

Conclusion. The teaching of CAS is a necessity for archivists to stay relevant and responsive to the changing landscape. We offer CAS graduate curriculum learning guidelines, ensuring that archives remain accessible, trustworthy, and reflective of our evolving society.

Introduction

Computing technology has been transforming the nature of archival work for decades. Initial transformations began with the rise of born-digital records in archives following the widespread adoption of computers and digital technology starting in the latter half of the 20th century and continuing apace even now. Additionally, the widespread adoption of computers and digital technology in the latter half of the 20th century opened the door for archivists to apply new computational practices and tools to archival work. One of the most significant changes has been the digitization of archival records. More recently, the application of artificial intelligence, for example, to the preservation, metadata generation, and identification of personal information has been transforming archival work (see, e.g., National Archives, 2016; Jaillant, 2022; Baron, Sayed & Oard, 2023). The rise of the internet, the web, and online databases have also made born-digital and digitized archival materials more accessible to a global audience. Researchers and the general public can now access archival collections from anywhere with an internet connection, reducing the need for in-person visits to physical archives. Additionally, advanced computing technology and techniques such as text mining, natural language processing, network analysis, and geospatial analysis have allowed researchers and archivists to perform data analysis and visualization on large sets of archival data to help uncover patterns, trends, and insights within collections, offering new perspectives and insights for researchers (Graham et al., 2016; Milligan, 2019). The many novel affordances brought about by the introduction of computing technology have also introduced novel challenges for archivists, however. Understanding the nature of new forms of born-digital records, such as those created using artificial intelligence or blockchain technology, remains an open challenge, as does how best to manage and preserve all types of born-digital records. In addition, the public and researchers are increasingly wishing to engage in new ways of researching archival materials that are potentially disruptive to traditional archival theories and practices (Marciano et al., 2019b). Archivists must now navigate new complexities related to the entrenched issues of digital privacy, copyright, and authenticity of records.

It is against this background of the transformative effects of computing technology on archival practice and theory, that the designation of computational archival science (CAS) was devised in 2016 by the two authors of this paper with five additional co-authors (Mark Hedges from King's College London, William Underwood formerly from GeorgiaTech, Michael Kurtz formerly from the U.S. National Archives, Mark Conrad formerly from NARA, and Maria Esteva, from the U. Texas Austin). A Foundational paper was drafted and published in 2018 (Marciano, 2018a), making a case for a new trans-discipline for CAS. In 2018, Nathaniel Payne, at the time a doctoral student of Victoria Lemieux, further proposed a revised definition of CAS as follows (Payne, 2018):

> *a transdisciplinary field grounded in archival, information, and computational science that is concerned with the application of computational methods and resources, design patterns, sociotechnical constructs, and human-technology interaction, to large-scale (big data) records/archives processing, analysis, storage, long-term preservation, and access problems, with the aim of improving and optimizing efficiency, authenticity, truthfulness, provenance, productivity, computation, information structure and design, precision, and human technology interaction in support of acquisition, appraisal, arrangement and description, preservation, communication, transmission, analysis, and access decisions.*

This chapter explores the kinds of CAS educational efforts and pedagogies that are emerging as part of archival science responding to the latest algorithmic and computational developments.

Context and development of computational archival science

The field of CAS has over 170 papers (see: Compendium of Core CAS Papers with 4 tabs:

https://docs.google.com/spreadsheets/d/1oCVCWpik_zjdilih9iXh2KITZ7LIOVIt/edit?usp=sharing&ouid=105652788602997060030&rtpof=true&sd=true) to date.

To promote the development of CAS, nine major CAS workshops were held as part of the international IEEE Big Data Conference with 119 papers presented, resulting in new CAS pedagogies (https://ai-collaboratory.net/cas/). CAS contributions alone represent five continents and 26 countries.

We exclude the growing number of papers that address the major themes and objectives of CAS that do not explicitly identify as CAS papers; these include books and papers generated by the AEOLIAN network (AEOLIAN network: https://www.aeolian-network.net/) and the InterPARES AI project (InterPARES Trust AI: https://interparestrustai.org/), to name only two examples. Combined with explicitly CAS papers, contributions from these projects send a strong signal that the archival field is taking a computational turn, albeit one with its own unique progressive and human-centred approach, as we will come to later.

One of the most important lessons learned has been the significance of integrating computational thinking (CT) concepts into archival science, which parallels the case for the inclusion of CT in mathematics and K-12 science classrooms. We will next describe a body of work on this topic that started in 2018. The emergence of Computational Archival Science parallels that of Computational Social Science, Computational Biology, and more recently Computational Journalism, with a specific focus on archival processing and theory.

(Underwood et al., 2018) and (Marciano et al., 2018b) are the first papers where working with LIS students, we introduced a CT framework organized as a taxonomy of 22 CT practices organized around four categories (see Figure 1). We showed how to map these practices through an archival case study involving World War II Japanese American Incarceration Camps focusing on automating the detection of personally identifiable information or PII.

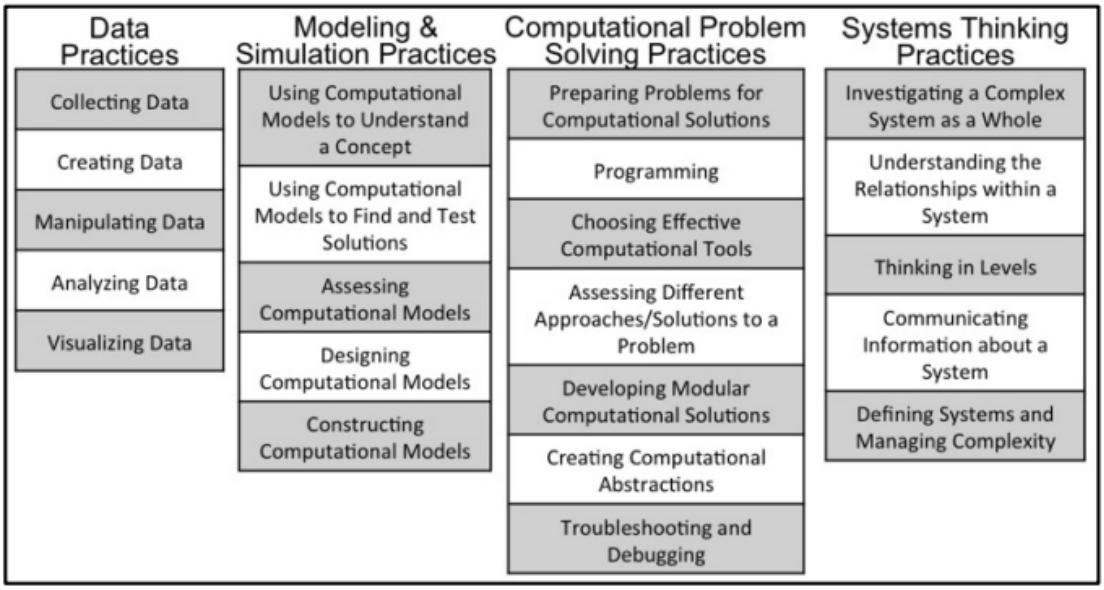


Figure 1. Computational thinking practices taxonomy

A subsequent larger study (Marciano et al., 2019a), involving 20 students (undergrad and grad), mapped five case studies to the CT taxonomy: (1) detecting personally identifiable information, (2) developing name registries, (3) integrating vital records, (4) designing controlled vocabularies, (5) mapping events and people, and (6) connecting events and people through networks. See Figure 2.

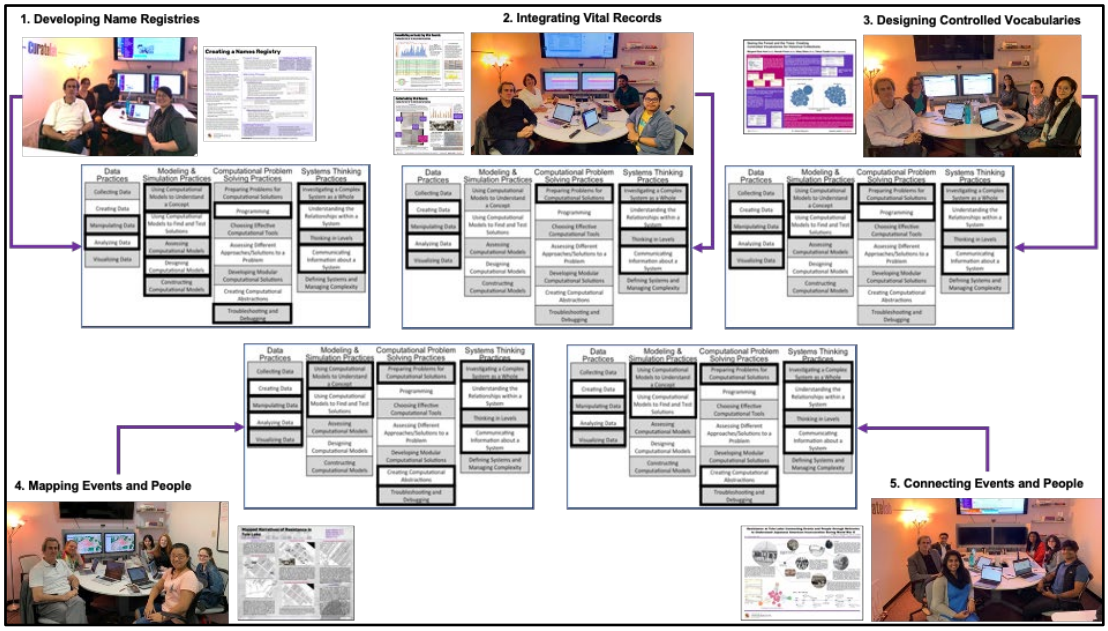


Figure 2. The CT taxonomy in practice with MLIS students

The value of mapping diverse CAS case studies into a common CT framework is clear as it provides unity and coherency in identifying the underlying computational skills and techniques that are relevant to archival work and a common language for teaching CAS to students.

Furthermore, what we were able to establish is that all 22 CT practices can be mapped into archival science case studies (Underwood et al., 2019a; Underwood, 2019b), demonstrating the generalizability of this approach. It is possible that in the future, in mapping in the other direction (i.e., from new computational archival projects to the 22 CT practices), some case studies may challenge the completeness of this original taxonomy. Furthermore, Buchanan et al., 2022 demonstrate experimentation with CT in archival science courses across four campuses and over 3 semesters of networked instruction.

As we write, the focus, goals, and disciplinary approaches within CAS remain unsettled and continue to evolve. Ongoing research into, analysis, and discussion of the focus, goals, and techniques of CAS as an emerging discipline will continue to play an important role in defining the competencies appropriate for students of CAS and the pedagogical content and techniques that best deliver those competencies, which we discuss in the next section.

Introducing CAS into graduate archival training: case studies

University of Maryland

In 2015, one year before the start of the CAS initiative, Richard Marciano and Michael Kurtz launched two pre-CAS pedagogical initiatives at the U. Maryland, the Digital Curation Innovation Center (DCIC) (https://ai-collaboratory.net/wp-content/uploads/2023/12/Brochure_2018-2019.pdf) and the Digital Curation for Information Professionals (DCIP) (https://info.umd.edu/academics/certificates-non-degree-study/digital-curation-for-information-professionals-certificate/).

The DCIC aimed to create a hands-on lab culture for LIS students, promoting interdisciplinary projects that explored the integration of archival research data and technology to generate new forms of analysis for historical research engagement. Projects were structured around social justice and infrastructure themes (https://ai-collaboratory.net/wp-content/uploads/2023/12/Project-Booklet_2018-2019.pdf). Student-led '*datathons*' included an

8-week UMD Digital Curation sprint on Japanese-American WWII Camps (https://ai-collaboratory.net/projects/ct-ja_ww2_camps/digital-curation-students-and-filmmaker-event/) and a 2-day Maryland State Archives (MSA) datathon on Legacy of Slavery (https://ai-collaboratory.net/projects/ct-los/student-led-datathon-at-the-maryland-state-archives/).
The DCIP was the U. Maryland iSchool's first professional certificate program designed for individuals currently in the library, archives, or digital curation field, or who were planning to enter it but wish to enhance their digital skills. The program (2015-2024), through three online courses over 9 months, took students from introductory through advanced digital curation lessons while focusing on practical applications and a final hands-on capstone project. Examples of projects include 2021 (https://ai-collaboratory.net/2021/08/23/dcip-cohort-2021-presents-capstone-projects/), 2022 (https://ai-collaboratory.net/2022/09/07/2022-dcip-cohort-presents-their-capstone-projects/), and 2023 (https://ai-collaboratory.net/2023/08/14/2023-dcip-cohort-presents-their-capstone-projects/), and 2024 (https://ai-collaboratory.net/2024/09/03/2024-dcip-cohort-presents-their-capstone-projects/). In 2020, with the COVID pandemic, Marciano and Lemieux joined forces with colleagues and pivoted to an online virtual research network, the Advanced Information Collaboratory (AIC) (https://ai-collaboratory.net/). The focus is on: (1) exploring the opportunities and challenges of "disruptive technologies" for archives and records management (including digital curation, ML, AI), (2) leveraging the latest technologies to unlock the hidden information in massive stores of records, (3) pursuing multidisciplinary collaborations to share relevant knowledge across domains, (4) training current and future generations of information professionals to think computationally and rapidly adapt new technologies to meet their increasingly large and complex workloads, and (5) promoting ethical information access and use.

Computational activities are now being incorporated into regularly scheduled LIS graduate courses. Our goal is to modernize archival and library education and contribute to the development of faculty and library "digital leaders." Examples include:

- INST747: **research in advanced digital curation** (see Jun. 18, 2022, student project https://ai-collaboratory.net/2022/06/18/june-18-2022-measuring-the-impact-of-urban-renewal/)
- INST604: **introduction to archives and digital curation** (see Dec. 5, 2023, final project https://ai-collaboratory.net/2023/12/05/dec-5-archival-processing-showcase/)
- INST742: **implementing digital curation** (see May 6, 2021, datathon (https://ai-collaboratory.net/2021/05/06/may-6-2021_datathon/), May 5, 2022, showcase (https://ai-collaboratory.net/2022/05/09/may-5-2022-digital-curation-project-showcase/), May 11, 2023, datathon (https://ai-collaboratory.net/2023/05/11/may-11-2023-computational-storytelling-datathon/), and May 9, 2024, datathon (https://ai-collaboratory.net/2024/05/08/may-9-2024-computational-storytelling-datathon/).
- **Spatial, graph, GenAI & LLM analysis**: In the summer of 2024, we introduced new MLIS courses on these topics applied to archives (https://ai-collaboratory.net/2024/07/09/summer2024/). See Figure 3.



Figure 3. A knowledge graph for urban renewal in Asheville, NC (*Author*: *Nick de Raet*)

Computational exposure was also incorporated into undergraduate Information Science courses. Examples include:

- INST341: **introduction to digital curation** (see Dec. 9, 2021 (https://ai-collaboratory.net/2021/12/14/dec-9-2021-digital-curation-showcase/)).

Incorporating computational research and training has led to media outcomes for students and portfolio projects they can incorporate into their CV, such as the followings:

- **Terp magazine** Winter 2023 – "*truth in exile*" (https://terp.umd.edu/truth-in-exile), with a project featured on '*NHK World Japan television*' (https://ischool.umd.edu/news/aic-umd-student-project-featured-on-japanese-television/).
- Public access portal that documents the hidden legacy of urban renewal in Asheville, NC (https://ischool.umd.edu/news/unearthing-the-hidden-legacy-of-urban-renewal-ashevilles-southside-neighborhood-reveals-intergenerational-impact/), and March 20, 2023, public testimony to the '*Asheville Community Reparations Commission*', Mayor, and City Council (https://www.youtube.com/watch?v=HO8vPhU0gLs).

Increasingly, we are exposing students (LIS grad and undergrad) to data science and archives through the creation of Jupyter Notebooks. See student feedback (Piety, 2023), with presentations and notebooks (https://ai-collaboratory.net/2023/05/11/may-11-2023-computational-storytelling-datathon/). The ultimate goal is to contribute to the development of faculty and library digital leaders. In addition to preparing our students to meet these challenges through research-oriented seminars, we are building an online repository called CASES (Computational Archival Research Educational System) involving CAS practices that address practical digital records management and archival problems (https://cases.umd.edu/). We are exploring how computational thinking practices might be introduced to graduate students in the core Archival

Studies curriculum. The goal is to enable a collaborative network of educators and practitioners who can learn from one another through the sharing and dissemination of computational case studies and lesson plans.

University of British Columbia

Teaching in CAS began at the University of British Columbia (UBC) in a course developed by Victoria Lemieux in 2011 on Information Visualization and Visual Analytics. The course, which began as a highly experimental topics course in its early days and continues to this day, now has a permanent place in UBC's Masters of Archival Studies (MAS) and Master of Library and Information Science (MLIS) programs. Since the takeaways from this early CAS work for archival education have already been discussed (Marciano et al., 2018a), here it is sufficient to remark that the content of this course was one of the first archival courses to expose students to the application of machine learning, artificial intelligence, and data visualization to archival work and research, Through the subsequent writing of Devon Mordell, who was a student in one of the offerings of the course, it also was able to seed the formulation of a new '*archives-as-data*' paradigm (Mordell, 2019).

Experimentation with CAS curriculum and pedagogy continued at UBC with the introduction of a summer institute on blockchain technology in 2017 and the formal introduction of a Blockchain Graduate Pathway in 2019. The program has been recognized by Coindesk as number one in Canada for blockchain education (Coindesk, 2023). In its current formulation, the blockchain graduate pathway aims to provide students with the knowledge and skills to contribute to the advancement of research and development in blockchain technology while at the same time engaging in some of the world's most complex sociotechnical issues. The multidisciplinary blockchain graduate pathway comprises a 12-credit non-degree training program that augments existing master's and PhD programs at UBC. Students in the pathway take courses on '*issues and perspectives on blockchain technology*'; technical '*foundations in blockchain and distributed ledger technology*'; blockchain and distributed ledger technology research; and '*blockchain for information professionals*', which delves into various records and information-related aspects of blockchain technology, such as their nature as records and record keeping systems, archival theoretic critiques of blockchain solution design, and compliance with regulations, including privacy, data protection and copyright.

In addition to completing coursework, students complete a research-based industry internship and prepare a major research paper or thesis based on the research conducted for the internship. The following papers represent examples of outcomes of this work: (Lemieux, Hofman, Batista, Joo 2019; Kang 2021; Lemieux, Voskobojnikov & Kang 2021; Suleman 2022; Suleman & Lemieux, 2023).

MAS and MLIS students who have completed this program have gained knowledge and experience equipping them to fill a range of archival, records management, and data and information management roles, as attested to by this former student:

> *I first heard of the blockchain @ UBC program in the iSchool newsletter. Having had no prior understanding of blockchain beyond knowing it had something to do with Bitcoin, I was interested in knowing more. The program provided me with an interdisciplinary knowledge of blockchain technology and the countless industries and use cases it can be applied to. Through the course work, the Summer Institute, and an internship with the Land Title and Survey Authority of British Columbia, I was able to apply theoretical concepts in the real world. While I currently do not work within the blockchain space, the transferable skills I gained through the program allowed me to become an information specialist at a technology company a few months prior to graduation.* (See: UBC iSchool, blockchain graduate pathway: https://ischool.ubc.ca/graduate/areas-of-focus/blockchainubc/).

The final example in this case study of integrating CAS into the curriculum for students of archives is a recently redesigned offering of a first-year core graduate course on archives on technology (ARST 500). The course aims to provide knowledge of the role of technology in archival work, at the theoretical and pragmatic level. Specific learning outcomes of this course are as follows:

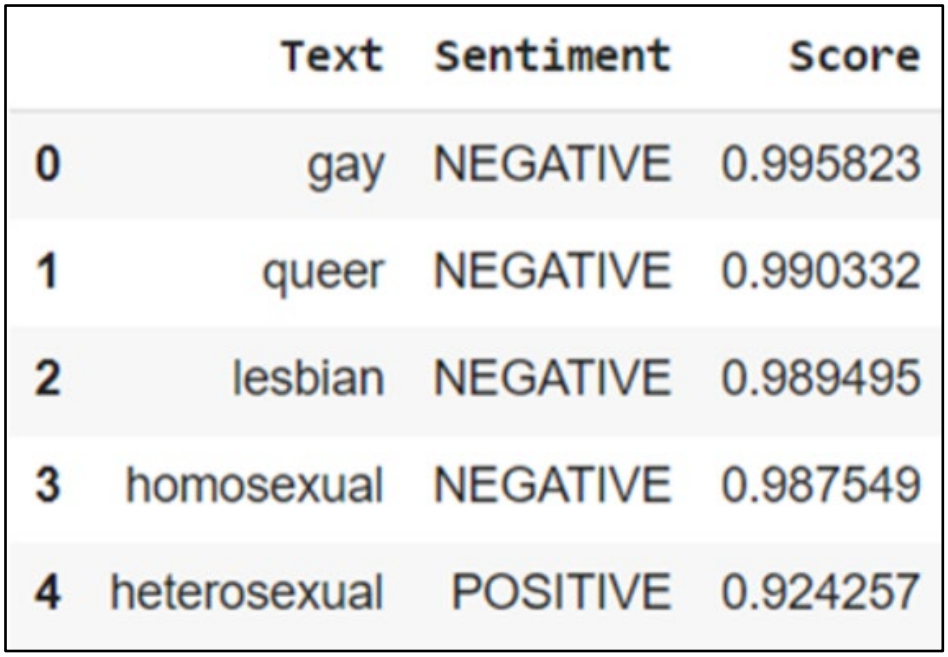
- Discuss the information and communication technologies in use in archives work and the affordances, both positive and negative, of these technologies
- Critically evaluate the differences between archival and computational theories and methodologies
- Apply knowledge of technology in support of archival work (e.g., set up a database and a website, use natural language processing, machine learning, and data visualization to analyse and make available information about archival materials
- Critically reflect upon the application of computational techniques and tools to archival documents or for archival work

Themes of archives as data; comparative notions of integrity and authenticity; knowledge representation; power and positionality; and privacy are also explored in this course.

One of the primary methods by which the learning outcomes are achieved is by assigning students a series of labs designed to help them explore the application of technology in support of archival work, and the implications of doing so in concrete and practical terms. In the 2023 offering of the course, groups of students were asked to conduct Natural Language Processing and Sentiment Analysis on supplied sample text extracted from a 1920 newspaper. To complete the lab, they used Google Colabs, a free Jupyter Notebook environment that requires no setup and runs entirely in the cloud, which eliminated the need for students to download software or set up a virtual environment to run the lab. Google Colab was used to run the Transformers Library by Hugging Face using the DistilBERT model. Transformers offers an open-source library that provides state-of-the-art machine learning models, primarily for tasks in natural language processing (NLP). As students were not expected to have prior programming experience, and the pedagogical purpose of the lab was not to teach them programming but rather to expose them to processes of datafication of archives, how natural language processing works, and the implications of using such tools in relation to data/records integrity and bias, they were given a '*recipe*' containing the script in Python to run the sentiment classifier rather than being asked to write the code themselves.

To stimulate students' critical thinking on the application of sentiment analysis to archival material, the lab was paired with readings, including (Mordel, 2019) and (Jo & Gebru, 2020), which students were asked to connect back with their experience of conducting the lab.

Several groups went beyond the specific assigned task to conduct further experiments using different sentiment classifiers and text. For example, one student group ran an experiment using the SpaCy sentiment classifier on Chinese text, generally observing that the results were consistent with what the students would have expected the classification of the words to be. Another group ran the DistilBERT model over specific words relating to sexuality (see Figure 4), concluding that the results showed a '*homophobic*' bias, and that the classifier was unable '*to understand the multiplicity of meanings [of] a word…which is quite problematic*'. This example of experiencing various flavours of sentiment analysis serves the purpose of developing critical thinking skills that show the strengths and weaknesses of computational techniques.



	Text	Sentiment	Score
0	gay	NEGATIVE	0.995823
1	queer	NEGATIVE	0.990332
2	lesbian	NEGATIVE	0.989495
3	homosexual	NEGATIVE	0.987549
4	heterosexual	POSITIVE	0.924257

Figure 4. Sentiment analysis for sexuality applied to single-word tokens

Building and sustaining educator networks

Successful examples of emerging educator networks can be found in Buchanan et al. 2023), where LIS students developed archival case studies with Jupyter Notebooks that document the International Research Portal for Records Related to Nazi-Era Cultural Property (IRP2), and a selection of unclassified catalogue entries about digitized nuclear science reports. Archival data visualizations are shown to promote outreach and access. Other examples include the LEADING network (LIS education and data science integrated network group) focused on preparing a cohort of LIS doctoral students and early to mid-career librarians for data science endeavours (Greenberg et al., 2023).

Beyond these emerging educator networks, to support the education of archival practitioners who are competent in CAS, it is necessary for there to be willing learners and archival educators to teach. This has been a challenge, as noted in a recent presentation made by Anne Gilliland on the theme of '*preparing archivists in computational thinking and innovative technologies*' at the international academic conference on digital intelligence: empowering the modernization of archival work, hosted by the school of history and culture, Shandong University, China in October of 2023,

> *It has been quite difficult to even develop the teachers of archivists, the professors who teach them, to develop those skills or even the awareness of where computational processes would be appropriate to use in archival work, archival work, as opposed to digital humanities, work or STEM. and I think it has been equally difficult to persuade students that this is not just an important part of archival work but an essential part of archival work.*

Our observation is that archival educators, by and large, lack the technical knowledge and skills to effectively teach CAS. Moreover, there is seldom institutional support to help them acquire the necessary knowledge and skills, as discussed in a recent paper by Lise Jaillant (2024). It takes incredible dedication and effort for someone '*classically trained*' in archival studies to acquire the technical capability to teach CAS. Equally, it takes training for anyone '*classically trained*' in technical disciplines, such as computer science or engineering, to acquire archival knowledge and

skills to teach CAS. The CAS educational challenge lies not only in the absence of technical knowledge and skills, however, but also in a lack of technical curricular content and pedagogical knowledge of how to teach such material to learners with very different levels of technical knowledge, skill and aptitudes.

Perhaps a greater challenge is motivation, though: while the transformation in archival work wrought by computation should be motivation enough for archival educators to '*grasp the nettle*', as Gilliland observes, this does not always happen. Similarly, more technical disciplines may lack interest in working with cultural materials and fail to appreciate the many other applications of archival theory (e.g., identifying deep fakes, to name but one '*au courant*' area).

To address the shortage of CAS educators and the many challenges of providing CAS education, and to stimulate a meaningful dialogue among educators of diverse disciplinary backgrounds with a view to developing CAS as a novel trans-discipline, we are now engaging in building a TALENT Network for the Training of Archival & Library Educators with iNnovative Technologies). See: *https://ai-collaboratory.net/projects/talent-network/*. TALENT brings together experts from across the United States (including archivists, librarians, Library and Information Science educators, historians, learning scientists, cognitive scientists, computer scientists, and software engineers) to create a durable, diverse, and multidisciplinary national community focused on developing digital expertise and leadership skills among archival and library educators. The network focuses on four major objectives: (1) doubling the educator network to include diverse educators in multidisciplinary iSchools that have a focus on adjacent disciplines (e.g., computing, engineering, education, and data science) where there are substantial needs for archivists and librarians with digital and computational skills, (2) engaging HBCU students (Historically Black Colleges and Universities), (3) conducting curriculum development through experts from a Learning Sciences Network, and (4) addressing the social and ethical concerns that arise from computational and algorithmic thinking. See Figure 5.

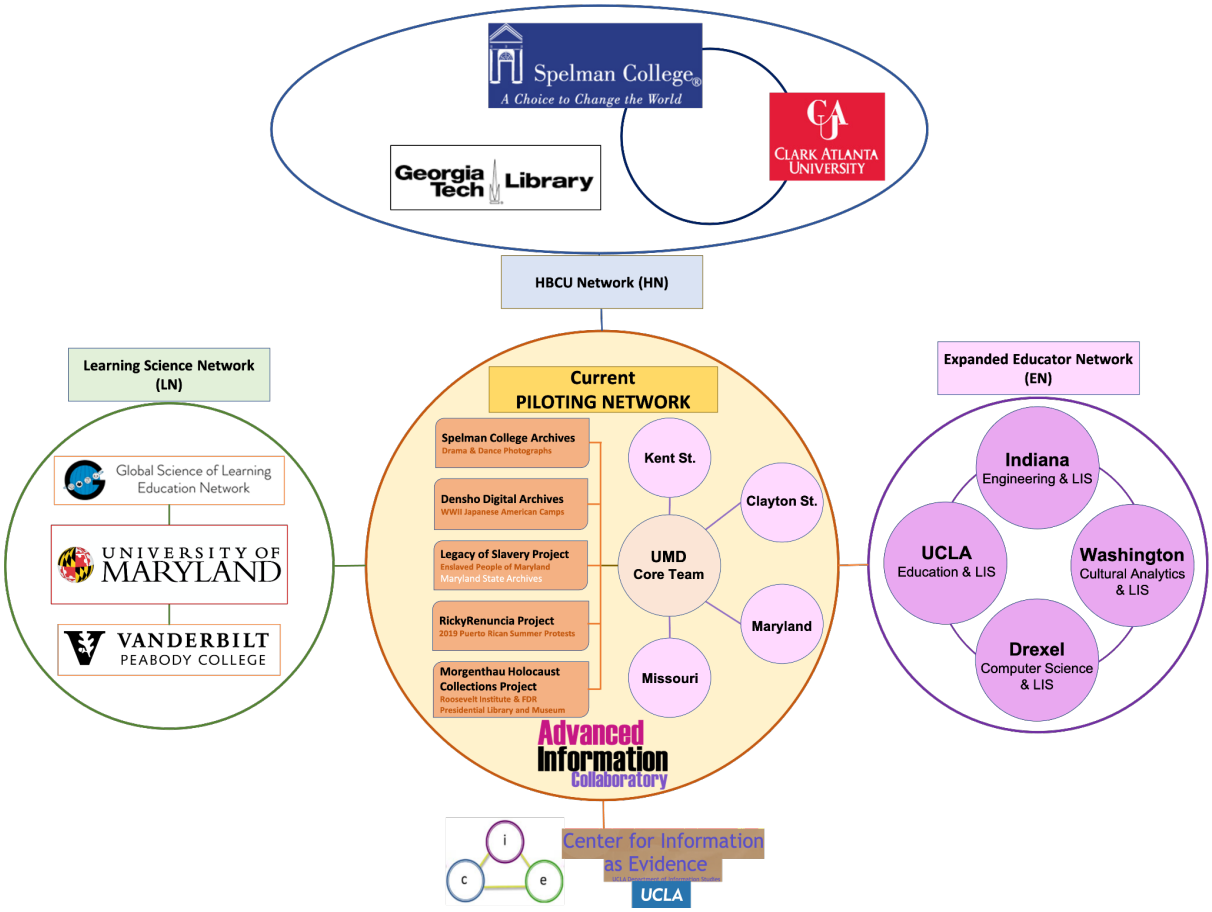


Figure 5. The TALENT network to broaden the teaching of CAS

All of the efforts to build and sustain CAS educator networks are for naught, of course, unless, as already mentioned, there are willing students. We note that in recent archival discourse, technology has often been associated with colonialism and other negatively perceived '*isms*', as this statement exemplifies:

> *in 2011, the Pluralizing the Archival Curriculum Group (PACG) of the Archival Education and Research Institute noted that "Archival studies education programs are conceptualized in strikingly similar ways worldwide, largely because of the overarching bureaucratically- and legally-centred paradigms developed and exported from Europe through colonialism, evangelism, mercantilism, and technological developments, and later codified through national and international standards and terminologies* (PACG, 2011).

Jaillant (2024) makes a similar observation about negative perceptions of technology. While efforts in pluralizing the archival curriculum are indisputably enriching, it is worth considering the potentially '*chilling*' effect such perspectives may have on students' willingness to learn about CAS and develop the technical skills that, as Gilliland (2023) and Jaillant (2024) observe, are not merely important but, nowadays, an essential part of archival work. This begs the question of how best to bring pluralizing perspectives into dialogue with CAS perspectives. Certainly, educational offerings focusing on the ethical aspects of applying technology to archival work, as discussed above, and those that exemplify how CAS techniques can be applied to advance social justice themes, help to achieve this end, but more effort is needed. The desire to advance this effort further is what, in part, motivated us to write this chapter, and also partly shaped our thinking on future directions for CAS, as we discuss in the following section.

Future directions

In articulating a new '*archives-as-data*' paradigm, Devon Mordell (2018), points to some avenues for expansion of the CAS construct. She refers, for example, to the '*invisibilisation*' of archives that can hide '*the human decision-making within the often-opaque interfaces of computational tools*'. Her reflection suggests an opportunity in CAS to methodically dissect and understand the intricate, interwoven layers of technical frameworks. This approach serves two purposes: firstly, to aid archival studies students in comprehending both the theoretical and tangible components of computing infrastructures, and secondly, to critically examine the discursive aspects of data as well as the inherent power dynamics and cultural positionality manifest within computational infrastructure and the logics of computing. By preparing students to think critically about these aspects, the approach helps to create a generation of archival professionals who are more likely to avoid falling into the trap of believing that technology is the solution to meeting every archival challenge and who can also anticipate, identify, and address ethical challenges and societal concerns in the application of technology in their work.

Fearing '*reactionary tendencies*' in CAS, Mordell further counsels the archival profession to ensure that a social justice critique is maintained within CAS. Madelynn Dickerson and Audra Eagle Yun's 2023 CAS workshop paper on '*critical community-centeredness: ethical considerations for computational archival studies*' signals how CAS might move in the direction of Mordell's recommendation. At the same time, the many papers highlighting the application of technology to advance social justice goals, such as Richard Marciano's work on the '*legacy of slavery*' and '*racial reparations for urban renewal*' illustrate how CAS techniques can uncover important insights to support processes that identify and help deal with the injustices of the past.

To further illustrate the impact of CAS on supporting past injustices, the Asheville Community Reparations Commission met on June 17, 2024, voting in support of a 10-page reparations request to the city amounting to $148K to families harmed by the displacement caused by urban renewal. A preliminary list of the names of individuals & businesses in the Southside neighborhood impacted and eligible for cash payments was compiled by students. See: https://docs.google.com/spreadsheets/d/110xsc3XXbm9Wa7klEJ75vnKw_b_BdimY/edit?gid=1080328214#gid=1080328214. This data was compiled using computational approaches as demonstrated in the resources listed in section 3.a. (public access portal, and public testimony links).

There also is room to use novel technologies to support archival educators in teaching CAS techniques, given the challenges of acquiring CAS knowledge and skills. In a recent CAS study, GPT-4's knowledge in some areas of archival practice, and its ability to think computationally about archival tasks are investigated (Underwood & Gage, 2023, Underwood & Gage, 2024). It is demonstrated that GPT-4 shows an understanding of 17 of the 22 CT practices.

> *These results support the possibility of ChatGPT+ being a partner to educators and students of archival studies in learning computational methods and tools applied to digital archival practice is a possibility. Similarly, archivists, records managers, and digital curators might be better able to apply computational methods and tools to their tasks with the aid of a ChatGPT+ agent.*

Ethical and trustworthy AI considerations should of course be part of this exploration, and indeed are a necessary component in developing computational archival pedagogies.

Finally, we think the time has come to move towards guidelines for graduate-level education in CAS. These guidelines should lay out the curricular content, pedagogy, learning outcomes, and graduate competencies to be developed. As a first effort, we propose the following set of competencies divided into three main areas of focus: technical, records/data, and social, based on

the development of CAS to date and following Lemieux and Feng's (2021) '*three-layer model*' of system design:

CAS graduate competencies:

Technical

1. Graduates possess technical knowledge and understanding enabling them to apply computational thinking, tools and methodologies to effectively analyze and preserve digital records and support archival work. Specifically, graduates have the ability to:

    1.1 apply knowledge of technologies to real world record creation and recordkeeping problems and situations

    1.2 demonstrate an understanding and skill in '*archives-as-data*' practices, such a collecting data, creating data, manipulating data, analyzing data, and visualizing data

    1.3 demonstrate an understanding and skill in how to apply modeling and simulation practices to archival work, such as using modeling to understand a concept, using computational models to find and test solutions, assessing computational models, designing computational models, and constructing computational models.

    1.4 within diverse contexts, demonstrate an understanding and skill in computational problem solving, including appreciating when it is appropriate to apply particular computational tools and techniques to specific archival tasks, preparing archival problems for computational solutions, programming, choosing effective computational tools, assessing different computational tools, developing modular computational solutions, creating computational abstractions, and troubleshooting and debugging.

    1.5 demonstrate the ability to think in terms of systems and apply systems thinking practices in archival work, including investigating a complex system as a whole, understanding the relationships within a system, thinking in levels, communicating information about a system, and defining systems and managing complexity.

Records/data

2. Graduates are able to apply the foundational theories, methodologies, and techniques of computational archival science to understand the effects of computing tools and methods on digital records and recordkeeping. Specifically, graduates have the ability to:

    2.1 demonstrate an ability to understand and apply foundational computational archival science theories to gain insights into the nature and evolving forms of digital records and recordkeeping.

    2.2 understand and critically evaluate the application of computational techniques and tools, such as those used in processes of '*datafication*', in relation to their impact upon the trustworthiness of records, such as their accuracy, reliability and integrity.

    2.3 understand and critically evaluate the impact of applying computational techniques and tools in relation to the selection and acquisition of records for preservation.

    2.4 understand and critically evaluate the impact of applying computational techniques and tools to sensitive information and its consequences for privacy.

2.5 demonstrate an understanding and ability to conduct risk assessments of threats and vulnerabilities to archival cyberinfrastructures and records and develop risk mitigation plans and strategies.

Social

3. Graduates are able to appreciate foundational theories, methodologies, and techniques of computational archival science to understand the effects of computing tools and methods on society. Specifically, graduates have the ability to:

    3.1 demonstrate an ability to critically examine the discursive aspects of data as well as the inherent power dynamics and cultural positionality manifest within computational infrastructure and the logics of computing.

    3.2 assess the social impact and usability of emerging technologies for archival purposes.

    3.3 draw upon knowledge of law and ethics in relation to all functions and activities related to applying computing technology to records and archives.

    3.4 engage critically and creatively with emerging professional and societal issues, such as bias in training data, arising from the application of computing technology to archival work

    3.5 advocate for the responsible and ethical application of computing techniques and tools in archival work.

We hope that programs offering archival education can use these proposed competencies to develop the programmatic CAS learning outcomes and curricular content that best speaks to their students and specific contexts. As more archival programs engage in teaching CAS, we believe that it would be beneficial for there to be a period of broader discussion facilitated by archival professional associations about the content of CAS graduate competencies and how best to align, or integrate, them with existing guidelines on graduate archival education.

A final point about pedagogy: our experience indicates that CAS requires students to '*get their hands dirty*' with the technology to truly understand it. This implies that students should learn by doing lab-based in-class activities and group or individual assignments that help to build their experience and skills in working with and applying technology to a range of archival tasks. We would add that this does not require archival educational programs to have access to, or create, advanced computing facilities. We have managed to teach CAS even using only free online tools and libraries.

Conclusion

Computing technology has revolutionized archival work, starting with the adoption of digital records and evolving with emerging technologies like AI and blockchain. This transformation includes digitization for easier access and preservation, applying advanced computational methods for data analysis, and tackling challenges in managing and preserving digital records. The rise of Computational Archival Science reflects these changes, focusing on computational tools and methods to improve archival practices and address emerging challenges in digital recordkeeping. Computational Archival Science (CAS) is still nascent and evolving, but it emphasizes integrating computing with archival science while advocating transdisciplinary collaboration. It also adopts hands-on learning, aiming to develop competencies across technical, record-keeping, and societal dimensions in archival education and the application of these competencies in archival work. As the case studies featured in this chapter illustrate, it is possible to teach CAS both to students with arts backgrounds and engineering backgrounds; in fact, to advance the development of CAS, pedagogy must be designed to support both groups. Additionally,

CAS does not have to be at odds with sensitivity to recent archival scholarship emphasizing postcolonialism and progressive themes – as the many CAS papers focusing on ethical and social justice aspects of the intersection of computing and archival work demonstrate, invalidating what. we referred to in the abstract as a false dichotomy.

The teaching of CAS is not merely an option but a necessity for the archival profession to stay relevant and responsive to the changing landscape. It will equip professionals with the tools and perspectives needed to navigate the complexities of digital recordkeeping, ensuring that archives remain accessible, trustworthy, and reflective of our evolving society. This commitment to evolving education in CAS highlights its significance in shaping a future where archives and the archival perspective are more than applicable to the past, but dynamic resources and ways of understanding and engaging with our present world.

Acknowledgments

On the U.S. side, we wish to acknowledge the support of three IMLS (Institute of Museum and Library Services) grants: (1) '*computational thinking to library and archival science education and research, CT-LASER*' (RE-73-18-0105-18 (https://www.imls.gov/grants/awarded/re-73-18-0105-18)) (2018-2020: https://ai-collaboratory.net/projects/ct-laser/final_report/), (2) '*piloting an online collaborative network for integrating CT into library and archival education and practice, piloting network*' (RE-246334-OLS-20 (https://www.imls.gov/grants/awarded/re-246334-ols-20)) (2020-2024: https://ai-collaboratory.net/projects/piloting-network/), and (3) '*training of archival & library educators with iNnovative technologies, TALENT network*' (RE252287-OLS-22 (https://www.imls.gov/grants/awarded/re-252287-ols-22)) (2022:2025: https://ai-collaboratory.net/projects/talent-network/). This work benefited from the contributions of Dr. Michael J. Kurtz, RIP Dec. 17, 2022, who was an inspiring mentor, collaborator, and dear friend of many in this CAS research community.

On the Canadian side, we acknowledge the support of '*Boeing Canada*' in developing training for archival and LIS studies students in visualization and visual analytics and Canada's '*Natural Science and Engineering Research Council*' funding for a Collaborative Research and Education Training Experience (CREATE) grant (CREATE 528125-1 (https://blockchain.ubc.ca/)) to establish the blockchain graduate pathway.

About the Authors

Dr Victoria Lemieux is a Professor of Archival Science at the School of Information, The University of British Columbia. She is also the Co-Lead of Blockchain@UBC research cluster. Her research interests focus on risk to the availability of trustworthy records, in particular in blockchain record-keeping systems, and how these risks impact upon transparency, financial stability, public accountability, and human rights. She can be contacted at vlemieux@ubc.ca

Dr Richard Marciano is a Professor at the University of Maryland in the College of Information. He is also the Director of the Advanced Information Collaboratory. His research interests focus on opportunities and challenges of disruptive technologies for archives and records management (including CAS – Computational Archival Science, AI, ML, Digital Curation, etc.), while promoting ethical information access and use. He can be contacted at marciano@umd.edu

References

Baron, J.R., Sayed, M.F. and Oard, D.W. (2022). Providing more efficient access to government records: A use case involving application of machine learning to improve FOIA Review for the

deliberative process privilege. *Journal on Computing and Cultural Heritage* (JOCCH) 15(1), 1–19. https://dl.acm.org/doi/10.1145/3481045.

Buchanan, S. A., Gracy, K. F., Kitchens, J. & Marciano R. (2022). Computational Thinking Integration into Archival Educators' Networked Instruction. In 2022 IEEE International Conference on Big Data, 7th CAS Workshop. https://ai-collaboratory.net/wp-content/uploads/2022/12/Sarah_Buchanan.pdf.

Buchanan, S. A., Wachtel J. L., & Stevenson J. A. (2023). Accelerating Precision Research and Resolution Through Computational Archival Science Pedagogy. In 2023 IEEE International Conference on Big Data, 8th CAS Workshop. https://ai-collaboratory.net/wp-content/uploads/2023/11/S01204_4947.pdf.

Coindesk (2023, May 11). Top Blockchain University: University of British Columbia. *Coindesk. https://www.coindesk.com/learn/top-blockchain-university-university-of-british-columbia/*

Gilliland, A. (2023). Preparing Archivists in Computational Thinking and Innovative Technologies. International Academic Conference on Digital Intelligence: Empowering the Modernization of Archival Work. School of History and Culture, Shandong University, China. https://www.youtube.com/watch?v=ZdLJHQLbR4k.

Graham, S., Milligan, I., Weingart, S. B., & Martin, K. (2016). *Exploring big historical data: the historian's macroscope*.

Greenberg, J. et al. (2023). IMLS-funded LEADING Project. https://mrc.cci.drexel.edu/leading/.

Jaillant, L. (2022). *Archives, access and artificial intelligence: working with born-digital and digitized archival collections*. Bielefeld University Press.

Jaillant, L. & Aske, K. (2024). Are Users of Digital Archives Ready for the AI Era? Obstacles to the Application of Computational Research Methods and New Opportunities. *Journal on Computing and Cultural Heritage* (JOCCH) 16(4), Article 87. https://doi.org/10.1145/3631125

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 306-316). ACM Digital Library.

Lemieux, V. L., & Feng, C. (2021). Conclusion: Theorizing from Multidisciplinary Perspectives on the Design of Blockchain and Distributed Ledger Systems (Part 2). In *Building Decentralized Trust: Multidisciplinary Perspectives on the Design of Blockchains and Distributed Ledgers* (pp. 129-163.). Springer International Publishing.

Lemieux, V., Hofman, D., Batista, D., & Joo, A. (2019). Blockchain technology & recordkeeping. *ARMA International Educational Foundation*. https://armaedfoundation.org/wp-content/uploads/2021/06/AIEF-Research-Paper-Blockchain-Technology-Recordkeeping.pdf

Lemieux, V., Voskobojnikov, A., & Kang, M. (2021, July). Addressing audit and accountability issues in self-sovereign identity blockchain systems using archival science principles. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 1210-1216). IEEE.

Ludäscher, B. (2016). A brief tour through provenance in scientific workflows and databases. In *Building trust in information: Perspectives on the frontiers of provenance* (pp. 103-126. Springer International Publishing.

Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M. & Conrad, M. (2018a). Archival records and training in the Age of Big Data. In J. Percell, L. C. Sarin , P. T. Jaeger , J. C. Bertot (Eds.), *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education* (Advances in Librarianship), Volume 44B (pp.179-199). Emerald Publishing

Limited. https://ai-collaboratory.net/wp-content/uploads/2020/10/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf

Marciano, R., Underwood, W., Hanaee, M., Mullane, C., Singh, A., Tethong, Z. (2018b). Automating Detection of PII in WWII Camp Records. 2018 IEEE International Conference on Big Data, 3rd CAS Workshop. https://ai-collaboratory.net/wp-content/uploads/2020/03/2.Marciano.pdf

Marciano, R. Aggarat, S., Frisch, H., Hunt, M.R., Jain, K., Kocienda, G., Krauss, H., Liu, C., McKinley, M., Mir, D., Mullane, C., Patterson, E., Pradhan, D., Santos, J., Schams, B., Shiue, H.S.Y., Silva, A.J., Suri, M., Turabi, T., Vaselli, M., & Xu, J. (2019a). Reframing Digital Curation Practices through a Computational Thinking Framework. In 2019 IEEE International Conference on Big Data, 4th CAS Workshop. https://ai-collaboratory.net/wp-content/uploads/2020/04/ReframingDC-UsingCT_final.pdf.

Marciano, R., Lemieux, V, Hedges, M., Tomiura, Y., Katuu, S., Greenberg, J., Underwood, W., Fenlon, K., Kriesberg, A., Kendig, M., Jansen, G., Piety, P., Weintrop, D., & Kurtz, M. (2019b). Establishing an International Computational Network for Librarians and Archivists. R., V. Lemieux, V., M. Hedges, Tomiura, Y., Shadrack, K., Greenberg, J., Underwood, W., Fenlon, K., Kriesberg, A., Kendig, M. In iConference 2019 Blue Sky Papers series, IDEALS Institutional Repository. http://hdl.handle.net/2142/103139.

Marciano, R. (2021). AFTERWORD: Towards a new Discipline of Computational Archival Science (CAS). In Jaillant, L (Ed.). *Access and Artificial Intelligence: Working with Born-Digital and Digitised Archival Collections*, Bielefeld University Press. https://www.transcript-open.de/doi/10.14361/9783839455845-009?html#read-container.

Milligan, I. (2019). *History in the age of abundance?: How the web is transforming historical research*. McGill-Queen's University Press.

Mordell, D. (2019). Critical questions for archives as (big) data. *Archivaria*, 87(87), 140-161. https://archivaria.ca/index.php/archivaria/article/view/13673.

National Archives (2016). The Application of Technology-assisted Review to Born-digital Records Transfer, Inquiries and Beyond. http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf

PACG (2011) Educating for the Archival Multiverse. *The American Archivist* 74, pp.69-101.

Payne, N. (2018). Stirring The Cauldron: Redefining Computational Archival Science (CAS) for The Big Data Domain. In 2018 IEEE International Conference on Big Data, 3rd CAS Workshop, https://ai-collaboratory.net/wp-content/uploads/2020/03/4.Payne_.pdf

Piety, P., Conrad, M., Marciano, R., Cornfield, I., Dallimore, E., Fettig, R., Hansen, E., Kemp, H., Turabi, T. (2023). Teaching and Learning with Archival Materials through the Development of Interactive Computational Notebooks. In Denison, V., Kearns, S. K., Leimkuehler, R., and Rogova, I. (Eds.), *2024 Archives and Primary Source Handbook*. New Prairie Press. https://ai-collaboratory.net/wp-content/uploads/2023/10/Piety_Conrad_Marciano_et_al-FINAL.pdf.

Suleman, Z. (2022). *Learning to trust: exploring the relationship between user engagement and perceptions of trustworthiness in self-sovereign blockchain systems* (Master's dissertation, University of British Columbia). https://open.library.ubc.ca/soa/circle/collections/ubctheses/24/items/1.0417415?o=1

Suleman, Z. J., & Lemieux, V. L. (2023). Learning to Trust: Exploring the Relationship between Trust and User Experience in Blockchain Systems. In *Blockchain in Healthcare: Analysis, Design and Implementation* (pp. 119-144). Springer International Publishing.

Underwood, W., Weintrop, D., Michael Kurtz, M., Marciano, R. (2018). Introducing Computational Thinking into Archival Science Education. In IEEE International Conference on Big Data, 3rd CAS Workshop. https://ai-collaboratory.net/wp-content/uploads/2020/03/1.Underwood.pdf.

Underwood, W. (2019a). Computational Thinking in Archival Science Research. Presentation at the Developing a Computational Framework for Library and Archival Education Workshop, iSchool Conference. https://ai-collaboratory.net/wp-content/uploads/2020/04/Underwood_CompThinkInArchResearch.pdf, with final report at: https://ai-collaboratory.net/wp-content/uploads/2020/11/Final_Report_r.pdf

Underwood, W. Marciano, R. (2019b). Computational Thinking in Archival Science Research and Education. In 2019 IEEE International Conference on Big Data, 4th CAS Workshop. https://ai-collaboratory.net/wp-content/uploads/2021/03/Underwood.pdf.

Underwood, W. Gage J. (2023). Can GPT-4 Think Computationally about Digital Archival Practices? In 2023 IEEE International Conference on Big Data, 8th CAS Workshop. https://ai-collaboratory.net/wp-content/uploads/2023/11/S01213_7999.pdf.

Underwood, W. Gage J. (2024). Training in Computational Archival Science: Do CAS Educational Frameworks meet Professional Expectations? In 2024 IEEE International Conference on Big Data, 9th CAS Workshop. https://ai-collaboratory.net/wp-content/uploads/2024/11/S01206_4692.pdf.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Gendered recordkeeping practices in marginalised communities in Bangladesh

Viviane Frings-Hessami

DOI: https://doi.org/10.47989/ir30iConf47344

Abstract

Introduction. Little is known about gendered recordkeeping practices in marginalised communities in developing countries and about how these practices have been affected by improvement in literacy in those communities.

Method. This paper reports the results of semi-structured interviews with 20 women and 17 men in two remote areas of Bangladesh about their recordkeeping practices.

Results. This paper shows that the female interviewees tend to preserve information more frequently than the male interviewees and that women are assuming more responsibilities in relation to keeping the important records of their family. The greater role assumed by women can be related to a rapid improvement in women's literacy in the past 30 years. However, comments made by interviewees show that this greater responsibility attributed to the female members of the households is not a reflection of women's empowerment. At the contrary, it is often an additional burden imposed on them since the women write down what their husbands expect them to write and keep the records that their husbands ask them to keep.

Conclusion. This study highlights the importance of doing more research on gendered recordkeeping practices and on their connections with women's empowerment in marginalised communities.

Introduction

Little is known about gendered recordkeeping practices in marginalised communities in developing countries and about how these practices have been affected by improvements in literacy in those communities. Recordkeeping is not a neutral activity, it is contingent on social and cultural norms and beliefs, including gender roles and expectations. Records are essential enablers of economic and social activities. In marginalised communities, for whom accessing information is difficult, capturing and preserving information that has been accessed so that it will be easy to re-access it if needed in the future can be beneficial. Due to improvements in women's literacy, women can play a greater role in regards to family recordkeeping than they did in previous generations.

This paper reports the results of semi-structured interviews with 20 women and 17 men in two remote areas of Bangladesh about their recordkeeping practices. It starts with a brief literature review on gendered recordkeeping, a discussion of literacy statistics in the areas where the research was conducted and an explanation of the methodology followed. Then, the findings section shows that the female interviewees tend to preserve information more frequently than the male interviewees and that women are assuming more responsibilities in relation to keeping the important family records. However, the author argues that this greater responsibility attributed to the female members of the households is not a reflection of women's empowerment. At the contrary, comments made by the interviewees indicate that it is an additional burden imposed on them by their husbands. The paper concludes by stressing the importance of doing more research on gendered recordkeeping practices and on their connections with women's empowerment and the empowerment of marginalised communities in developing countries.

Literature review

The creation and use of records are impacted by the socio-economic conditions in which they are created and by the values, ideas and aspirations of their creators and users. Social circumstances influence what information individuals have access to, how they interpret it and how they are able to capture it, preserve it and use it (Nesmith, 2006). Ketelaar (2005) wrote that recordkeeping is not a neutral activity, but contingent on social and cultural norms and beliefs. In many societies, women are considered to be the guardians of the family memory (Ketelaar, 2005), '*the customary nurturers and keepers of family tradition*' (Davison, 2000, p. 81). Research conducted in Iceland showed that women were traditionally regarded as the keepers of the family's records, mementos and memorabilia because they were responsible for the home while the men were working outside (Tucker and Bogadóttir, 2008). However, very little research has been conducted on recordkeeping practices in developing countries.

Ketelaar wrote in 2005 that gendered recordkeeping, in particular, had not yet been recognised sufficiently by scholars and practitioners in archival studies. Gender influences what records individuals have access to, what information they are interested in, how they can access it and use it, and what methods they can use to capture and preserve important information and records. Rutkair (2011) argued that gender constitutes an important element of the societal provenance of records. In the past ten years, several authors have applied feminist theory to rethinking the construction and use of archives (Dever, 2018; Caswell & Cifor, 2016; Caswell, 2021; Lapp, 2023). However, these studies have focused on records intended to be preserved as archives because of their cultural value rather than on records needed to meet the pressing socio-economic needs of their creators in the short term, and overwhelmingly on Western urban settings. No research has been done about gendered recordkeeping in developing countries where women are still subject to many socio-cultural restrictions and the ways they interact with information are greatly restricted by the socio-cultural expectations imparted on to them as daughters, wives and mothers.

Bangladesh is a developing country affected by frequent flooding and natural disasters where 45% of the population make their living from agriculture. It is a context where capturing information previously accessed and keeping it for future reference can save arduous journeys to nearby towns in search of information from government officials. It is also a country where women living in rural areas are still restricted by socio-cultural and religious norms that make it difficult for them to leave their homes and their villages to access information that would enable them to take advantage of economic opportunities (Lewis, 2011; Frings-Hessami & Oliver, 2023).

Little research has been done on recordkeeping and information preservation practices in rural areas of Bangladesh. However, two previous studies conducted by the author in the same areas as the research for this paper provided the motivation for this research and offer interesting points of comparison. The first study conducted in the southern sub-district of Shyamnagar found that villagers who had limited access to smartphones and to the internet were very conscious of the fragility of digital technologies and used paper notebooks as a form of backup for information that they found on the internet (Frings-Hessami, 2023). That study, which was based on a relatively small sample of interviews, also found that both men and women wrote information in notebooks and that they wrote the same types of information, such as the names of medicines, fertilisers and pesticides recommended by the local vets or found online, or new farming techniques learnt during training sessions. Although the study found differences between the ways men and women accessed information due to the fact that men were more likely to have a smartphone that they could use to access information, it did not find any significant differences in the ways men and women kept records and preserved information. The second study conducted in the sub-district of Dimla in the north of Bangladesh found that all the female participants asserted that in their families, women were in charge of keeping the family's important records, such as land titles, birth certificates, diplomas and identity cards (Frings-Hessami, 2024). The women in this study also used notebooks to write information that they found important. The two studies showed that more research should be done to understand the potential differences between the recordkeeping practices of men and women in marginalised communities of Bangladesh.

Literacy levels in Bangladesh

Literacy levels in Bangladesh have improved dramatically in the past 30 years from 38.9% in 1991 to 76.56% in 2021 for men and from 25.5% in 1991 to 72.82% in 2021 for women (Bangladesh Bureau of Statistics, 2014, p. xi; Bangladesh Bureau of Statistics, 2022, p. 17). The gap between male literacy and female literacy has closed down. However, inequalities remain between rural and urban population.

According to the 2011 census, in the sub-district of Dimla, only 43.6% of the rural men and 38.8% of rural women were literate (Bangladesh Bureau of Statistics, 2014, p. 449). The census defined literacy as 'the ability to write a letter in any language' (Bangladesh Bureau of Statistics, 2014, p. 6). In terms of educational achievements, the 2011 census reported that 41.28% of rural men and 47.36% of rural women in Dimla had not completed any schooling and 21.15% of the men and 19.6% of the women had completed four years or less of schooling (Bangladesh Bureau of Statistics, 2015a, pp. 209, 226).

Literacy rates are higher in the sub-district of Shyamnagar. According to the 2011 census, 52.6% of the rural men and 43% of the rural women were literate (Bangladesh Bureau of Statistics, 2014, p. 358), and 65.11% of rural men and 63.71% of rural women had more than 4 years or less of schooling (Bangladesh Bureau of Statistics, 2015b, pp. 273, 278).

In those two sub-districts, the improvement in literacy is reflected in the differences between the literacy levels of adults aged 20-24 and those aged 40-44. The 2011 census reported that among the rural population aged 40-44 in Dimla, 14.67% of women and 29.5% of the men were literate, whereas 59.77% of the women and 61.13% of the men aged 20-24 were literate (Bangladesh Bureau

of Statistics 2015a, pp. 154-155). In Shyamnagar, 46.12% of the rural men and 22.62% of the rural women aged 40-44 were literate, compared to 72.1% of the men and 69.31% of the women age 20-24 (Bangladesh Bureau of Statistics 2015b, p. 176).

Methodology

This paper reports the findings from semi-structured interviews conducted in January 2023 in two remote and marginalised areas of Bangladesh: in the sub-district of Dimla, district of Niphamari in the north of the country and in the sub-district of Shyamnagar, district of Satkhira in the south of the country. Twenty women and 17 men (10 women and 7 men in Dimla; 10 women and 10 men in Shyamnagar) were interviewed about the ways they access information and about their recordkeeping practices. All the interviewees, male and female, were involved in agricultural activities. In addition, 5 officers from local NGOs who were working on development projects in each area (3 in Dimla and 2 Shyamnagar, all male) were also interviewed.

The interviewees were selected by local NGO officers who were well acquainted with the communities so as to represent a variety of age groups. Tables 1 and 2 present the demographic characteristics of the community people who were interviewed. They show that the education level achieved by the participants is higher in Shyamnagar than in Dimla, which is in line with the general statistics for those two areas. The male interviewees are older than the female interviewees in both locations, which is a reflection of the greater availability of older men to attend an interview because younger men are more likely to be working away from home. All the interviewees from Dimla are Muslims, whereas the majority of the interviewees from Shyamnagar are Hindus, coming from predominantly Hindu villages.

Interview questions were prepared by the author who had previously conducted research in both areas, but the Bangladeshi research assistant who conducted the interviews in Bangla was advised to conduct the interviews in a semi-formal way and to follow up interesting comments with additional questions. The interviews were audio-recorded with consent from the participants, then translated into English by Bangladeshi research assistants before being analysed. The author read each interview transcripts several times and coded the emerging themes in an iterative way. Findings relating to information literacy, to information behaviour and to information preservation will be discussed in other papers. This paper focuses on an interesting, and unexpected theme that emerged while re-reading and analysing the transcripts, that of gendered recordkeeping and its connections with women's literacy and empowerment.

Participant code	Gender	Age	Religion	Education level
DW1	F	40	Muslim	Grade 4*
DW2	F	22	Muslim	Middle school certificate
DW3	F	22	Muslim	Grade 5
DW4	F	22	Muslim	Middle school certificate
DW5	F	21	Muslim	Middle school certificate
DW6	F	40	Muslim	Grade 5
DW7	F	40	Muslim	Grade 9
DW8	F	27	Muslim	Grade 8
DW9	F	35	Muslim	Grade 8
DW10	F	22	Muslim	Grade 8
DM1	M	33	Muslim	Grade 8
DM2	M	26	Muslim	Postgraduate
DM3	M	42	Muslim	Grade 5
DM4	M	32	Muslim	Middle school certificate
DM5	M	50	Muslim	Grade 8
DM6	M	50	Muslim	Grade 5
DM7	M	51	Muslim	Grade 8

Table 1. Demographic characteristics of community participants - Dimla.
* Grade 4 means 4 years of schooling, Grade 5: 5 years of schooling, etc

Participant code	Gender	Age	Religion	Education level
SW1	F	35	Muslim	Grade 4
SW2	F	34	Hindu	College
SW3	F	28	Hindu	Grade 10
SW4	F	30	Hindu	College
SW5	F	35	Hindu	College
SW6	F	38	Hindu	Postgraduate
SW7	F	30	Hindu	Grade 10
SW8	F	38	Muslim	Middle school certificate
SW9	F	27	Hindu	High school certificate
SW10	F	36	Hindu	Postgraduate student
SM1	M	39	Hindu	High school certificate
SM2	M	52	Muslim	Grade 7
SM3	M	33	Hindu	Middle school certificate
SM4	M	28	Hindu	College
SM5	M	23	Hindu	Postgraduate
SM6	M	?	Hindu	College student
SM7	M	43	Muslim	College
SM8	M	Over 60	Hindu	Grade 8
SM9	M	30	Muslim	Grade 8
SM10	M	Around 50	Hindu	Grade 8

Table 2. Demographic characteristics of community participants - Shyamnagar

Findings

Preserving important records

The interviews showed that in most families, women are now in charge of keeping the important records of the family. However, there are some significant differences between the two sub-districts in which interviews were conducted.

In the northern sub-district of Dimla, all the female interviewees reported that they were responsible for keeping the important records of their family (see Table 3). They explained that it was their responsibility because they were home all the time while their husbands were working outside. Most of the male interviewees (6 out of 7) also said that the female members of their family, usually their wives were responsible for keeping the important records, but some of them commented that they organised the records, then gave them to their wives to keep them safely. For example, DM5 asserted that he organised the important documents in files and explained to his wife how to keep them in order. Similarly, DM7 commented that '*I organised everything and give it to her [his wife] and tell her to keep all these things carefully*'. DM1 stressed that in case of flood, it was him who carried the records to safety over his head. DM6 commented that the practice used to be that the male elders were in charge of keeping the money and the important things, but that nowadays '*money and other important things are stored by the wives.*'

Participant code	Person responsible for looking after the family records
DW1	self (wife)
DW2	self (wife)
DW3	self (wife)
DW4	self (wife)
DW5	self (wife)
DW6	self (wife)
DW7	self (wife)
DW8	self (wife)
DW9	self (wife)
DW10	self (wife)
DM1	women
DM2	women
DM3	wife
DM4	self (husband)
DM5	wife
DM6	wife
DM7	wife

Table 3. Responsibility for important family records - Dimla

In the southern sub-district of Shyamnagar, most of the female interviewees (8 out of 10) also reported that they were responsible for keeping the important records of their family. Only one woman said that her husband was responsible for looking after all the records, while another woman said that her father-in-law kept the important records of the family, such as land records, but that she kept her identity documents and her birth certificate herself. However, most of the male interviewees declared that taking care of the important records was the responsibility of the men in their family. Five of them said that they took care of the records themselves, one said that his father did it, and two said that they did it conjointly with their wives, while only two of them said that their wives were responsible for taking care of the family records.

Participant code	Person responsible for looking after the family records
SW1	husband
SW2	self (wife)
SW3	self (wife) & father-in-law
SW4	self (wife)
SW5	self (wife) & husband
SW6	self (wife)
SW7	self (wife)
SW8	self (wife)
SW9	self (wife)
SW10	self (wife)
SM1	self (husband) & wife
SM2	self (husband)
SM3	self (husband)
SM4	self (husband)
SM5	father
SM6	self (husband) & wife
SM7	self (husband)
SM8	wife
SM9	self (husband)
SM10	wife

Table 4. Responsibility for important family records - Shyamnagar

Although the contrast between the answers of the male and female interviewees may be due to the small size of the sample, it could also be explained by the different ways in which the male interviewees and the female interviewees interpreted the question. The men's answers reflect their desire to show that they are in charge of their family's affairs and their understanding that although their wives are responsible for the day-to-day task of looking after the family records (e.g. have the key to the box in which the records are kept so that they can access their identity documents when their husbands are away), the men are responsible for making the decisions on what records to keep and how to keep them. The women's answers, on the other hand, reflect the actual responsibility that they assume on a day to day basis.

It is interesting to note that although most of the interviewees from Shyamnagar were Hindus (9 out of 10 of the women and 7 out of 10 of the men), the three Muslim men admitted that their wives were responsible for or shared the responsibility of looking after the family records, whereas 6 out of the 7 Hindu men said that it was the men's responsibility. However, given the small size of the sample, this potential correlation requires further research.

Preserving information in notebooks

The interviews showed that the practice of writing information in notebooks so that it can easily be accessed when required is common in the two areas of the study, but with some interesting differences between the two areas and between the practices of the male and female interviewees.

In Dimla, most of the female interviewees (7 out of 10) said that they wrote down in a notebook information that they found important, such as names of fertilisers and pesticides for their crops and of medicines for their animals. They also keep track of their income and expenses by writing them in a notebook. One young woman (DW5) also reported that she likes to write stories. Two of the three interviewees who did not write information were illiterate (DW1 and DW6). DW1 nevertheless understood the importance of writing down important information and regularly asked her daughter to write information, such as the names of medicines for their animals and

names of seeds, in a notebook for her. Only one interviewee (DW8) said that she rarely wrote information because she preferred to go and ask her experienced relatives who lived nearby when she needed information.

All the male interviewees from the same area, except one, said that they wrote information in a notebook, but most of them admitted that they did not write much. The interviewee who did not capture information (DM7) declared that he did not do it because when he needed advice, he just went and asked the agriculture extension officers for it.

Female and male interviewees agreed that women write more than men. The women attributed this to the fact that the men work outside the home for long period of time whereas the women stay at home. Some women mentioned that the men in their area were less literate than the women and one of them reported that her husband could not write. DW1 commented that: '*The women write more. The men in our area are less literate. Nowadays we are educating girls more*'.

However, the fact that the women write more is not just a reflection of the education achievements of younger women, it is actually the result of men's expectations that their wives will retain useful information, as well as keep track of all their income and expenses. DW7 commented that '*The men are mostly outdoors. So, we, women, tend to write [information] more, and then we give this to the men, and they use it*'. And DW10 explained that women write more than men because '*men always ask about records from the women*' and '*women can't remember all the time*'. The men from their area agreed with this explanation. For example, DM2 admitted that most of the writing was done by the women and that the men told them what to write: '*The advice comes from us. We told them to write it down. They preserve the information by writing it*'. However, one of the men (DM1) reported a more egalitarian approach to capturing information. He said that both he and his wife wrote information in notebooks: '*Whenever one of us gets the opportunity, we write*'.

The three NGO officers from Dimla who were interviewed agreed with the comments that women write more than men. According to one of them, in the houses where women are literate, they do most of the recordkeeping because men do not get much time. He commented that: '*So men usually tell women the price of things and ask them to write it, and this is how they keep the record of things.*' Another NGO officer concurred that women are more inclined to write things down than men, and that usually the males would tell the females: '*you write it down and keep it with you.*'

The interviewees revealed a bigger contrast in Shyamnagar where all the female interviewees said that they write down some type of information in notebooks, but few men did. Most of the women (8 out of 10) write down agricultural information, while two women do not write agricultural information, but write information that they find important such as important dates, or information that they find interesting. SW9 who is a postgraduate student as well as a farmer's wife commented that she wrote whatever she felt like, for example song lyrics.

On the other hand, most of the men from that sub-district declared that they did not write down agricultural information because they felt that they could memorise the information that they needed or that they could easily get information when they needed it by going to the local shops. Those who wrote down information said that they wrote very little and in an unorganised way on pieces of paper that they put in their pocket and may not find again when they need them. Several of them said that they wrote phone numbers in a small notepad and one of them also mentioned that he wrote the names of medicines for his family. An older man (SM8) commented that he used to write down information, but that because his eyesight was not good, he asked his children to do it for him. And one of the younger men (SM4) who did not write agricultural information reported that he was writing songs and lyrics because he was interested in making music. However, all the men said that they took charge of writing the income and expenses from their aquaculture activities.

Like in Dimla, the female interviewees commented that women write more than men. For example, SW1 declared that: '*the women in this area have some educational qualifications so they try to write down as much as possible*'; and she added that: '*men don't do this much*'. The two male NGO officers from Shyamnagar who were interviewed agreed that the tendency of writing agricultural information was more common among women than men. However, none of the men from that location admitted that the women wrote more than the men. Three of them asserted that their wives did not write information and SM7 said that his wife could not write. As discussed in the previous section, the discrepancy between the comments made by the men and those made by the women could be due to the small size of the sample or could be a reflection of the men's reluctance to admit that they are not in charge of all the affairs of their family. Given that the male NGO officers concurred with the views of the female interviewees, the latter explanation is more likely, although it is also possible that the men might not be aware of how much agricultural information their wives were preserving to support the activities, they were responsible for, such as growing vegetables and raising chickens and ducks near their homes.

Discussion

The interviews showed that in most families, women are now in charge of keeping the important records of the family. This is in contrast with the practice of previous generations when the male elders were assuming that responsibility as the head of the family. This new responsibility can be linked with improvements in female literacy in the past 30 years. In some families, the women have higher education levels than the men. In previous generations when women were not literate, they were not given the responsibility to look after the family records. However, the reason usually given for entrusting the women with that responsibility is that they are at home all the time while men work outside the home all day or may be away from the village for work for months at a time. Although the women like having access to the records they need for themselves and their children when their husbands are away, they also see this as one more responsibility for them.

Women also write down more information in notebooks than men, but several interviewees, male and female, asserted that this is because the men tell them to do it. In particular, husbands expect that their wives will account for every cent they spent. Therefore, the wives have to carefully write down all their expenses. In some cases, men also ask their wives to write down agricultural information. Like looking after the family records, this is one more chore that is being imposed onto the women.

However, writing down agricultural information in notebooks to preserve it for future reference is something that women often do of their own initiative because they understand the benefits of doing it. This is particularly useful in communities which do not have access to smartphones (Frings-Hessami, 2023) as is the case of the majority of the participants in this study since preserving information saves them the time and the cost of having to find the information again. And this is even more important for women who are restricted in their movements due to cultural expectations. Therefore, although looking after records and writing down the information that their husbands asked them to keep are additional chores for the women, writing down information that they believe is useful to them and their families, if they can spare the time to do it, can be seen as beneficial and liberating. Some young women also enjoy doing other types of writing in their spare time, such as keeping a journal or writing recipes.

The men, on the other hand, are more focused on keeping track of the income and expenses of their agricultural activities and less interested in writing agricultural information because they can more easily access information when they need it by going to the agricultural supplies shops or talking to agriculture extension officers. They delegate some of the day-to-day tasks of looking after the family records and keeping track of income and expenses when they work outside the home, but still see themselves as the persons in charge of those tasks in their households.

Conclusion

The findings from this study show that Bangladeshi rural women assume important responsibilities in relation to keeping the important records of their family and capture useful information that their family may need in the future. These new responsibilities can be related to literacy improvements. However, they do not constitute an indication of women's empowerment since those responsibilities are imposed on to them by their husbands who give them clear directions about the records to preserve and the way to preserve them. Therefore, the findings from this study show that there is a need for more research on gendered recordkeeping practices and on the connection between recordkeeping and women's empowerment in marginalised communities.

Acknowledgements

The research for this paper was supported by a Discovery Early Career Research Award (DECRA) Fellowship from the Australian Research Council (Project No DE210100012).

About the author

Viviane Frings-Hessami is a Senior Lecturer and Co-Director of the Information Empowered Communities Lab in the Faculty of Information Technology, Monash University, Australia. Between 2021 and 2023, she was the recipient of Discovery Early Career Research Award (DECRA) Fellowship from the Australian Research Council. She holds a PhD in Political Science and a Postgraduate Diploma in Information and Knowledge Management from Monash University and a MA in Southeast Asian Studies from the University of Kent at Canterbury. Her current research focuses on recordkeeping and information preservation in marginalised communities.

References

Bangladesh Bureau of Statistics. (2014). *Population & Housing Census 2011. National volume 2: Union Statistics*. Accessed 26 October 2023 at: https://web.archive.org/web/20170903181037/http://203.112.218.65:8008/WebTestApplication/userfiles/Image/National%20Reports/Union%20Statistics.pdf

Bangladesh Bureau of Statistics. (2015a). *Population & Housing Census 2011. Zila Report: Nilphamari*. Ministry of Planning, Statistics and Information Division. Accessed 26 October 2023 at: https://web.archive.org/web/20210825111738/http://203.112.218.65:8008/WebTestApplication/userfiles/Image/PopCenZilz2011/Zila_Nilphamari.pdf

Bangladesh Bureau of Statistics. (2015b). *Population & Housing Census 2011. Zila Report: Satkhira*. Ministry of Planning, Statistics and Information Division. Accessed 13 September 2024 at:http://203.112.218.65:8008/WebTestApplication/userfiles/Image/PopCenZilz2011/Zila-Satkhira.pdf

Bangladesh Bureau of Statistics. (2022). *Population and Housing Census 2022 Preliminary Report*. Ministry of Planning, Statistics and Information Division. Accessed 13 September 2024 at: https://sid.portal.gov.bd/sites/default/files/files/sid.portal.gov.bd/publications/01ad1ffe_cfef_4811_af97_594b6c64d7c3/PHC_Preliminary_Report_(English)_August_2022.pdf

Caswell, M. (2022). Dusting for fingerprints: Introducing feminist standpoint appraisal. Journal of Critical library and Information Studies, 3(2), https://journals.litwinbooks.com/index.php/jclis/article/view/113

Caswell, M., & Cifor, M. (2016). From human rights to feminist ethics: Radical empathy in the archives. Archivaria, 81, 23–43. https://archivaria.ca/index.php/archivaria/article/view/13557

Davison, G. (2000). *The use and abuse of Australian history*. Allen & Unwin.

Dever, M. (Ed.). (2018). *Archives and new modes of feminist research.* Routledge.

Frings-Hessami, V. (2023) Information sustainability in rural Bangladesh: the use of analogue and digital backups. In Sserwanga I, Goulding A, Moulayson-Sandy H, Du JT, Soares AL, Hessami V, Frank RD (eds) iConference 2023: Normality, Virtuality, Physicality, Inclusivity, LNCS 13971, pp. 66–78. https://doi.org/10.1007/978-3-031-28035-1_6

Frings-Hessami, V. (2024). Motivations for personal recordkeeping practices: the roles of personal factors, recordkeeping literacy and the affordances of records. Archival Science 24(1), 83-99. https://doi.org/10.1007/s10502-023-09431-9

Frings-Hessami, V. & Oliver, G. (2023). Accessing and preserving information: combining ICT4D and archival science to empower marginalized communities. Journal of the Association for Information Science and Technology, 74(12), 1350-1364. https://doi.org/10.1002/asi.24702

Ketelaar, E. (2005). Sharing: Collected memories in communities of records. Archives and Manuscripts, 33(1), 44–61.
https://publications.archivists.org.au/index.php/asa/article/view/9759

Lapp, J.M. (2023). 'The only way we knew how': provenancial fabulation in archives of feminist materials. Archival Science, 23(2), 117–136. https://doi.org/10.1007/s10502-021-09376-x

Lewis, D. (2011). Bangladesh: politics, economy and civil society. Cambridge University Press.

Nesmith, T. (2006). The concept of societal provenance and records of nineteenth century Aboriginal-European relations in Western Canada: Implications for archival theory and practice. Archival Science, 6(3-4), 351–360. https://doi.org/10.1007/s10502-007-9043-9

Rutkair, J. (2011). Adding gender to the archival contextual turn: The Rocky Mountain photographic records of Mary Schäffer Warren. MA thesis, Department of History, University of Manitoba, Winnipeg.

Tucker, S., & Bogadóttir, S. (2008). Gender, memory, and history: in one culture and across others. Journal of Archival Organization, 6(4), 288–310.
https://doi.org/10.1080/15332740802563161

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

The affective dimension of archival work: understanding the thoughts and feelings of archivists who documented the HIV/AIDS epidemic

Allan A. Martell, Travis L. Wagner, and Shannon Oltmann
DOI: https://doi.org/10.47989/ir30iConf47341

Abstract

Introduction. Affect is a vehicle of meaning-making that helps individuals adapt to stressful conditions. Affective experiences help meaning-making during life transitions and crises, but less is known about the affective experiences of information professionals, such as archivists.

Method. Employing a qualitative research approach, we conducted 25 semi-structured interviews with archivists working in LGBTQIA+ archives across the United States. We asked about participants' thoughts and feelings about their work documenting the HIV/AIDS epidemic in the United States.

Analysis. Our emerging findings are based on an analysis of the first 10 interviews of our dataset. Our data analysis relied on inductive and deductive coding, concept mapping, and memo writing.

Results. Archivists who documented this epidemic had to interact with vast amounts of information about loss and grief. Our participants experienced a physical toll due to this, reflected upon the impact of the epidemic on the LBTQIA+ community, and felt a range of negative and positive affects.

Conclusion. Archivists' affective experiences are more complex than just burnout. Our preliminary findings suggest that the affective experiences of archivists documenting trauma and crises may not only lead to burnout but also to experience appreciation for the role of solidarity and the positive impact of documentation on the LGBTIA+ community.

.

Introduction

Information has an affective dimension (Boehner et al., 2005; Clore et al., 2001). In fact, affects are a type of embodied cognition that matters to information science. *Affects* are defined as shifts in the psyche. When individuals attach meanings to these shifts with words, these are referred to as *feelings* (Brennan, 2003). As information sources, affects enable individuals to make meaning during life transitions such as when migrating (Allard & Caidi, 2018), undergoing treatment for chronic health conditions (Genuis & Bronstein, 2017), and transitioning genders (Huttunen & Kortelainen, 2021). Meaning-making is associated with greater subjective well-being, especially when individuals are faced with stressful work environments -- such as health-care workers (Krok et al., 2021) and older adults during the COVID-19 pandemic (Lind et al., 2021). Meaning-making has also been shown to increase adaptive responses among disaster survivors (Park, 2016). In other words, affect is a vehicle of meaning-making that enables individuals to increase their subjective well-being and adapt to stressful conditions. Exploring affective experiences is important to the field of information. How individuals interact with information can therefore impact their capacity to adapt to challenging environments and their sense of subjective well-being.

The archival profession is one of the disciplines within the information field where concern about the subjective well-being of information workers has been on the rise. Emerging work has strongly suggested that archivists working with records related to trauma and crises (e.g., extreme weather events, forced migration, armed conflicts) may undergo significant stress (Laurent & Hart, 2021). Tools are being developed to mitigate archivists' burnout (Crisis & Disaster Working Group, 2023; Laurent & Hart, 2021). These tools currently focus on delivering information to people affected by crises, such as protocols for working with victims of extreme weather events or people forcefully displaced due to armed conflicts. Notwithstanding the value of these materials, more work is needed to ensure information professionals, like archivists, have the necessary tools to process the significant stress of engaging in this type of work.

In this article, we report on a project that aims to bridge this gap. We look at the case study of the archivists who documented the HIV/AIDS epidemic in the U.S. The HIV/AIDS epidemic peaked in the 1980s and 1990s. In these twenty years, official estimates account for over eight hundred thousand people who lost their lives to the virus (Department of Health and Human Services, 2001). The LGBTQIA+ community was the most vulnerable, with incidence rates of contagion five to six times higher than the general population (Department of Health and Human Services, 2001). Those infected experienced social isolation and were ostracized by their relatives due to fears of contagion and of others learning about the sexual orientation of the infected person (Audet et al., 2013; Emlet, 2006). We posit that archivists who documented the HIV/AIDS epidemic had to interact with vast amounts of information about loss and grief. Therefore, their experiences are revealing of the affective journeys of information professionals handling information related to trauma and crises.

Methods

This article is part of a larger study about the documentation practices of LGBTQIA+ archives. In a prior publication, we discussed the value of community accountability in ensuring the long-term sustainability of these archives (Wagner et al., 2024). In this piece, we sought to answer the question: what are the affective experiences of the archivists who documented the HIV/AIDS epidemic? Following Brennan (2003), we studied the different aspects of affective experiences: physical sensations and feelings and the meanings participants derived from them.

Our data comes from conducting semi-structured interviews (Kvale, 1996; Rubin & Rubin, 2005; Seidman, 2006) with archivists working for LGBTQIA+ archival institutions. Data was collected between May and August of 2023. We compiled a list of 69 LGBTQIA+ archival organizations in the United States by conducting Google searches and based on our knowledge of well-known queer

archival institutions. We emailed all institutions in the list describing our project and calling for participants. Respondents were asked to complete a demographic survey and schedule a semi-structured interview. Interviews lasted between 60 and 90 minutes. Our interview protocol was divided into five segments:

1. *An introduction to the archive*
2. *The impact of the participants' identity to their work in an LGBTQIA+ archive*
3. *Archival policies for collection development*
4. *Accessioning and deaccessioning practices*
5. *Uses of their LGBTQIA+ archive*
6. *Advice for creating an LGBTQIA+ archive*
7. *Closing questions and debriefing.*

The second segment contained the questions specific to archivists' affective experiences:

1. *What thoughts and feelings did you experience when processing materials related to the AIDS epidemic?*
2. *How did you handle those thoughts and feelings? For instance, did you go to therapy, talk to someone, or felt at least the need to do so, or took a break from work?*

We analysed our interviews using an inductive and deductive approach divided into two phases (Charmaz, 2006). In the first phase, the three authors used conceptual codes drawing from their respective disciplines of information curation and social memory (first author), queer media archives and digital curation (second author), and information access and ethics (third author). In the second phase, the first author and their graduate student assistant drew from in vivo, process, and descriptive codes. We inferred codes from the data itself. In the second phase, we used axial coding to weave our codes into a consistent narrative of our data (Kelle, 2007). We also drew concept maps (Miles et al., 2014) to visualize the relationships between codes. The map served as the basis for the results section. Throughout the coding process, we wrote analytic memos (Emerson R. M., 2011; Lempert, 2007) to derive and interpret our findings and draw our conclusions. These memos served as the foundation for the present report.

Results

25 people from across 22 archival institutions participated in this project. At the time of this writing, we have completed the analysis of the first 10 interviews in our dataset. Though all our participants worked in institutions devoted to documenting the experiences of LGBTQIA+ people, not all of them were members of this community. Seven identified as queer, two as straight, and one chose not to disclose their sexual orientation. All participants expressed at least one emotion about their archival work. These emotions ranged broadly, from sadness and exhaustion to a sense of fulfilment.

Sensations

Some of the terms expressed by our participants to describe their emotions referred more to physical sensations, such as feeling '*drained*' (P5) and '*heavy*' (P10). When asked about the thoughts and feelings they experienced while documenting the HIV/AIDS epidemic, P5 noticed that:

> *I have to make sure that I am not getting overly invested in my work sometimes because it is emotionally draining. Just because a lot of it is just a lot of tragedy. So, trying to balance that is something that I definitely find different* (P05).

P05's remark that '*it is just a lot of tragedy*' has two meanings: first, that the epidemic impacted a heavier burden on the LGBTQIA+ community than the rest of the community. Second, that P05, in their role as an archivist, had to engage very closely, for long periods, with multiple materials related to the epidemic, including personal diaries, photographs, and other records of people who

were infected with and die from the virus. This closeness and repeated exposure to such information had a draining effect on P05.

Thoughts

Another participant, P02, referred to their experience not with a specific word but with a reflection on the impact of the epidemic on the LGBTQIA+ community. They recalled the conversations he had with his friends during the 1990s when they met regularly at a gay bar after work:

> *hearing some of the stories and thinking about like, the one person's dad who died alone in a small town, and, you know, there I'd had done some work in like the late 90s in the, in the Iron Range, and there was one couple and one partner had AIDS and the community and 100 like, rallied around any medical issues or anything like that. They've just rallied around and so thinking about, you know, what, what does that say about this particular person's family? You know, they're in terms of just isolation like, I think that was the biggest thing I thought about is like, how different it is just in certain camps (P02).*

P02's experience with epidemic-related records made him reflect upon how the epidemic had a magnifying effect on the isolation that many LGBTQIA+ individuals already experienced (Audet et al., 2013; Emlet, 2006). Importantly, P02 experience during the 1990s allowed them to appreciate the concrete actions of fellow community members to show solidarity with those infected, thereby mitigating the impact of social isolation on those living with the virus.

Negative feelings

Two participants reported challenging feelings, such as sadness and grief (P04, P05) and '*unprocessed trauma*' (P09). For P04 and P05, this emotional challenge stemmed from the taxing nature of archiving the epidemic, whereas for P09, the challenge stemmed from the limited resources to support donors:

> *the other thing too, I would say about queer collecting is there's a lot of unprocessed trauma. And we found out, maybe a little too late, that we needed to be able to offer some of our older donors' access to psychological services. And that's one of the, I guess that would be technically volunteer, we have a group of therapists who will volunteer with us in case we have a donor that is having a real difficult time processing. There's a lot of unprocessed trauma, especially around HIV, a lot of guilt. That's been hard too, it really has (P09).*

P09's experience highlights how archivists' emotional challenges are linked to handling records with difficult stories and to their personal connections to trauma survivors. In this sense, P09's claim that '*it's been really hard*' refers as much to himself as to patrons', therefore highlighting the caring nature of the archival profession (Agostinho, 2019). It appears that archivists experience duress when they perceive that they are unable to their caring duties towards patrons, which in this case is due to limited resources.

Positive feelings

In addition to these challenging feelings, whether experienced by our participants personally or noticed in their donors, some archivists also reported positive feelings, such as confidence (P01), empathy (P02), pride (P04), awe (P08), joy (P09) and fulfilment (P10). These feelings coexisted with the challenging ones mentioned earlier as can be seen in P10's case. P10 described themselves as belonging to an '*in-between*' generation, meaning that they could remember the impact of the HIV/AIDS epidemic in the second half of the 1990s. In that sense, they noticed feeling '*heavy*' when processing epidemic-related materials. Nonetheless, they also felt fulfilled:

> *I guess, like, doing the work is both sort of exhausting and taxing and hard, but it's also fulfilling. And so, it's like you get energy from doing it, it also saps your energy. And so, I don't know. It's sort of a balance. You come back to things, and the break helps you to be able to continue on. Certainly, seeing the results of the work, whether that's some scholar has published an article and they used our archives, and I can just be so proud of the fact that I personally was able to help them and that we had these materials* (P10).

P10's remark is important because it evinces a complex emotional experience where the challenging and rewarding aspects of their emotional experience exist in a relationship with each other. While acknowledging that processing epidemic-related records takes a toll, P10 also notices that undertaking this work has a positive social impact. As P10, notices that others use the records they have processed, P10 finds that their documentation efforts have a positive impact. To P10, this impact provides an emotional reward that makes their work worthwhile.

Discussion

The emerging patterns in our data analysis suggest that archivists who documented the HIV/AIDS epidemic experienced a physical toll, reflected upon the impact of the epidemic on the LBTQIA+ community, and felt a range of negative and positive affects. The physical sensations of heaviness and drain are more closely connected to stress than relaxation. This suggests a closer connection to negative affects, even if archivists did not explicitly name these using feeling words. In this sense, our findings mirror what prior studies about affect and archives have already shown: archivists documenting trauma and crises do experience their work as a burden (Laurent & Hart, 2021). However, our preliminary findings also reveal that such documentation may allow archivists to experience positive affect. Examples of these are witnessing the positive impact of solidarity among community members, which to some degree mitigates the isolation experienced by those living with HIV/AIDS and realizing the social relevance of the documentation being produced.

It appears that negative and positive feelings exist in tandem, informing each other. P10's remarks suggest that at least some participants found not just solace, but motivation to continue with their work when they perceived that the archival records they produced were embraced by others. Furthermore, positive feelings such as P04's, P08's, and P10's may indicate that archivists have positive associations with their documentation work. In particular, P10's experience tentatively suggests that these positive associations assist archivists navigate the burdensome aspects of their labor. Thus, documenting trauma and crises may not only be a source of burnout to archivists but also provide some degree of reward. This has practical implications for those engaged in trauma-informed archival practice (Laurent & Hart, 2021; Laurent & Wright, 2020; Sexton, 2025). Future interventions in this discipline can help archivists promote and appreciate community solidarity and the positive social impact of the documentation work in addition to the strategies already in place for mitigating the risks of trauma and alleviating its effects (Crisis & Disaster Working Group, 2023). Based on our results, we believe this strategy holds the potential to increase the subjective well-being of archivists who document trauma and crises.

Archivists need not only to receive resources that help them better assist victims but also resources to support themselves in this task, both physically and emotionally. Such resources can help archivists reduce the impact of negative affects and accrue the benefits of positive ones.

Acknowledgments

Many thanks to the archivists who participated in this project and our home institutions.

About the authors

Allan A. Martell is an Assistant Professor in the Department of Information and Library Science at Indiana University Bloomington. Martell explores how societies negotiate social memories of violence; the role of information curation in shaping such memories; and frameworks to promote more critical, nuanced memories. Martell received his PhD in information from the School of Information at the University of Michigan, an MS in digital media from the Georgia Institute of Technology, and a BA in social communications from the Central American University (El Salvador). Martell also worked as a postdoctoral researcher at Louisiana State University.

Travis L. Wagner is an Assistant Professor in the School of Information Sciences at the University of Illinois Urbana–Champaign. Wagner's research interests include critical information studies, queer archives, and LGBTQIA+ advocacy in sociotechnical systems. Their work investigates how LGBTQIA+ communities create identity in opposition to sociotechnical systems that characterize and limit those identities. Additionally, Wagner studies the unique relationships between obsolete archival mediums and queer counter-historical work across archival contexts. Wagner earned their PhD in information science from the University of South Carolina (USC) and served as a postdoctoral fellow at the Center for Archival Futures (CAFe) and the Recovering and Reusing Archival Data (RRAD) lab within the University of Maryland's College of Information Studies. They are the co-creator of the Queer Cola Oral History and Digital Archive.

Shannon M. Oltmann is an Associate Professor in the School of Information Science, College of Communication and Information, at the University of Kentucky. Oltmann's research focuses on issues of intellectual freedom, censorship, privacy, and information ethics. Her work is mostly qualitative, examining how and why civil liberties such as freedom of speech become curtailed or enhanced. Much of her work has centred on public libraries, but Oltmann has also studied government agencies, art galleries, social media platforms, public court records, and other types of libraries.

References

Agostinho, D. (2019). Archival encounters: rethinking access and care in digital colonial archives. Archival Science, 19(2), 141–165. https://doi.org/10.1007/s10502-019-09312-0

Allard, D., & Caidi, N. (2018). Imagining Winnipeg: The translocal meaning making of Filipino migrants to Canada. Journal of the Association for Information Science and Technology, 69(10), 1193–1204. https://doi.org/10.1002/asi.24038

Audet, C. M., McGowan, C. C., Wallston, K. A., & Kipp, A. M. (2013). Relationship between HIV Stigma and Self-Isolation among People Living with HIV in Tennessee. PLoS ONE, 8(8), e69564. https://doi.org/10.1371/journal.pone.0069564

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2005). Affect. Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility, 59–68. https://doi.org/10.1145/1094562.1094570

Brennan, T. (2003). The Transmission of Affect. Cornell University Press.

Charmaz, K. (2006). Constructing grounded theory: a practical guide through qualitative analysis (pp. xiii, 208 p.). SAGE. http://mirlyn.lib.umich.edu/Record/005210104

Clore, G. L., Gasper, K., & Garvin, E. (2001). Affect as Information. In Handbook of Affect and Social Cognition (pp. 68–93). Psychology Press. https://doi.org/10.4324/9781410606181-8

Crisis & Disaster Working Group. (2023, February). Documenting in Times of Crisis: A Resource Kit. Society of American Archivists. https://www2.archivists.org/advocacy/documenting-in-times-of-crisis-a-resource-kit

Department of Health and Human Services. (2001). HIV/AIDS Surveillance report: U.S. HIV and AIDS cases reported through December 2001. https://www.cdc.gov/hiv/pdf/statistics_2001_HIV_Surveillance_Report_vol_13_no2.pdf

Emerson R. M., F. R. I. & S. L. L. (2011). Writing ethnographic fieldnotes. University of Chicago Press.

Emlet, C. A. (2006). An Examination of the Social Networks and Social Isolation in Older and Younger Adults Living with HIV/AIDS. Health & Social Work, 31(4), 299–308. https://doi.org/10.1093/hsw/31.4.299

Genuis, S. K., & Bronstein, J. (2017). Looking for "normal": Sense making in the context of health disruption. Journal of the Association for Information Science and Technology, 68(3), 750–761. https://doi.org/10.1002/asi.23715

Huttunen, A., & Kortelainen, T. (2021). Meaning-making on gender: Deeply meaningful information in a significant life change among transgender people. Journal of the Association for Information Science and Technology, 72(7), 799–810. https://doi.org/10.1002/asi.24447

Kelle, U. (2007). The Development of Categories: Different Approaches in Grounded Theory. In A. Bryant & K. Charmaz (Eds.), The SAGE Handbook of Grounded Theory (pp. 191–213). Sage Publications.

Krok, D., Zarzycka, B., & Telka, E. (2021). Risk of Contracting COVID-19, Personal Resources and Subjective Well-Being among Healthcare Workers: The Mediating Role of Stress and Meaning-Making. Journal of Clinical Medicine, 10(1), 132. https://doi.org/10.3390/jcm10010132

Kvale, S. (1996). Interviews: an introduction to qualitative research interviewing. Sage Publications. https://mirlyn.lib.umich.edu/Record/003062848

Laurent, N., & Hart, M. (2021). Building a trauma-informed community of practice. Education for Information, 37(1), 27–32. https://doi.org/10.3233/EFI-190363

Laurent, N., & Wright, K. (2020). A trauma-informed approach to managing archives: a new online course. Archives and Manuscripts, 48(1), 80–87. https://doi.org/10.1080/01576895.2019.1705170

Lempert, L. B. (2007). Asking questions of the data: memo writing in the grounded theory tradition. In A. Bryant & K. Charmaz (Eds.), The SAGE Handbook of Grounded Theory (pp. 245–264). Sage Publications.

Lind, M., Bluck, S., & McAdams, D. P. (2021). More Vulnerable? The Life Story Approach Highlights Older People's Potential for Strength During the Pandemic. The Journals of Gerontology: Series B, 76(2), e45–e48. https://doi.org/10.1093/geronb/gbaa105

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). Qualitative data analysis: a methods sourcebook. SAGE. https://mirlyn.lib.umich.edu/Record/012444368

Park, C. L. (2016). Meaning Making in the Context of Disasters. Journal of Clinical Psychology, 72(12), 1234–1246. https://doi.org/10.1002/jclp.22270

Rubin, H. J., & Rubin, Irene. (2005). Qualitative interviewing: the art of hearing data (pp. ix, 291 p.). Sage Publications. http://mirlyn.lib.umich.edu/Record/004768705

Seidman, I. E. (2006). Interviewing as qualitative research: a guide for researchers in education and the social sciences (3rd ed.). Teachers College Press. http://mirlyn.lib.umich.edu/Record/002448412

Sexton, A. (2025). Introducing the legacies and trajectories of trauma to the archival field. Archival Science, 25(1), 3. https://doi.org/10.1007/s10502-024-09471-9

Wagner, T. L., Martell, A. A., & Oltmann, S. M. (2024). It's Not the Materials Themselves, It's the Attitude of the Donors": The Role of Community Accountability in the Sustainability of Queer Archives. Archivaria.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

An ensemble framework for sentiment-embedded event evolution in diaspora oral archives

Jing Zhou

DOI: https://doi.org/10.47989/ir30iConf47338

Abstract

Introduction. Diaspora oral archives should be displayed in the context of diversity and inclusiveness rather than being glossed over by the dominant, normative group so their voices need to be spread further.

Method. This paper proposes an ensemble framework for sentiment-embedded event evolution in diaspora oral archives. It contains a knowledge representation model, an event evolutionary graph, and an event extraction workflow to extract entities, events, and relationships.

Results. The South Asian Oral History Project is selected as the data source. The key events, entities, event types, sentiment and event relations are extracted with natural language processing techniques to construct sentiment-embedded event evolutionary graph. Based on this, the event evolution, spatio-temporal and spatio-sentiment patterns are analysed.

Conclusion. Such methods allow researchers and archivists to engage in research on machine-assisted oral archives to ensure reproducibility, reduce interpretative biases, and efficiently and swiftly amplify hidden voices of '*the other*'.

Introduction

Diasporas' voices and stories are ignored in the mainstream public consciousness. Oral archives may communicate hidden voices, revealing alternative narratives to those published in written histories (Brinkhurst, 2012). In terms of data types, oral archives are the audio/video recordings with field notes and transcriptions collections. These interviews can take various shapes including autobiographical narratives by the interviewee, such as lifestyle interviews, or interviews with semi-structured, open-ended questions following a research agenda, such as thematic interviews, or collective information sessions, where many persons participate in the conversation (Thomson, 1998). Hence, diasporas oral archives always need to be read in relation to broader, dislocated contexts. The contexts include an account of events, personal affection, etc. Nevertheless, on the one hand, similar oral archive practices concentrate more on the collection and preservation processes, neglecting semantic features and dynamic correlation in stories so events in these projects are barely connected by now. On the other hand, the participants-generated stories are unstructured textual data that are not predefined and indexed so it's hard to integrate or retrieve entities and events from them. This calls for new models and processes that enable intuitive access to event-related knowledge in diaspora oral archives. Hence, this study will focus on events and sentiments in diaspora oral archives, together with related personal narrative elements. These events and their relations, as well as entities, are vital components of diasporas' stories, constituting basic knowledge units of diaspora oral archives.

Diaspora oral archives are usually highly personal and emotive stories. The sentiment in archives creates a new legibility of the individual narratives (Roeschley and Kim, 2019; Jones, 2019). This study introduces the concept of "sentiment-embedded events" to describe events within these narratives that are closely tied to emotional responses, particularly those evoked by specific locations and contexts. Using a sentiment analysis framework that categorizes emotions into positive, neutral, and negative, this research adopts a coarse-grained approach to analyse sentiments associated with places and events. While this study does not aim to capture the finer nuances of affective experiences, it provides an overarching perspective on the emotional patterns linked to immigration activities and events.

An ensemble framework for sentiment-embedded event evolution is employed in this study, including a knowledge representation model, an event evolutionary graph (EEG) and automatic event extraction workflow by natural language processing (NLP). Precisely, this research exploits the sentiment-embedded event evolutionary graph (SEEGraph) to interpret and analyse these events and their relations, including the sentiment entailed in events. Besides, summarizing diaspora oral archives through manual annotation can be very tedious (or nearly impossible for long texts or large corpora). Many tasks have been successfully adapted to machine learning models in NLP that can help humans extract and summarize information from text automatically, as well as enhancing consistency and scalability in data analysis. Hence, events and their relations, and sentiments will be extracted through NLP in this study.

Motivated by these ideas, we hope to answer the following questions in our research:

– What knowledge representation model can capture sentiment-embedded events and their relationships in diaspora oral archives?

– How can machine learning models be integrated effectively to extract and enrich information in the proposed knowledge representation model?

– What are the meaningful patterns and insights of sentimental trajectories and event evolution in diaspora narratives?

Related works

Some researchers began to pay attention to knowledge representation of diaspora oral archives by text mining and NLP. These latest technologies create unique possibilities for the analysis of oral history interviews (Pessanha and Salah, 2021), such as identifying key topics within the histories, including events, and social or political issues (Rieping, 2022; Brown and Shackel ,2023).

Recently, sentiment in diaspora archives has drawn some researchers' attention. For instance, the First Days Project (Caswell and Mallick, 2014), Harvest Moon Oral History and the Flin Flon Heritage Project provide opportunities to acquire, describe, and preserve affective records that recount affective emotions (Grant, 2020). As for the automatic identification of complicated sentiment in diaspora archives, sentiment analysis is a relatively new approach for retrieving affective information in contexts of diaspora oral histories and interviews (Dominguès et al., 2019; Ozdemir and Bergler, 2015).

To further extract and store complete events and their relations, a conceptual model and SEEGraph are introduced in this study to standardize knowledge representation and enhance affective narration. The EEG originates from the concept of the event knowledge graph, which emphasizes the realization of goals such as events logic mining, and generalization. Event knowledge graphs have previously been employed in areas such as news articles (Rudnik et al., 2019), and contemporary and historical events (Gottschalk et al., 2023). The SEEGraph is proposed to provide a foundational descriptive framework for the event extraction results in NLP, while also incorporating sentiment within the events.

Research method

Knowledge representation model

The classes, properties, and hierarchical structures in the CIDOC CRM ontology are reused to construct the ontology of diaspora oral archives. CIDOC CRM is a high-level ontology that could play the role of the mediated schema in the information integration of cultural heritage data and their correlation with library and archive information (Bekiari et al., 2022). In CIDOC CRM classes like E53 Place, E2 Temporal Entity, E5 Event, E39 Actor, E55 Type are always connected and represented with an Event. The ontological framework of diaspora oral archives is shown in Figure 1. Entities in the ontology include Event, Event Type, Action, Participant, Time, Location, and Sentiment. Relations between Events are temporal relation and causal relation. As for relations between entities, this research reveals spatio-temporal relations and spatio-sentiment relations.

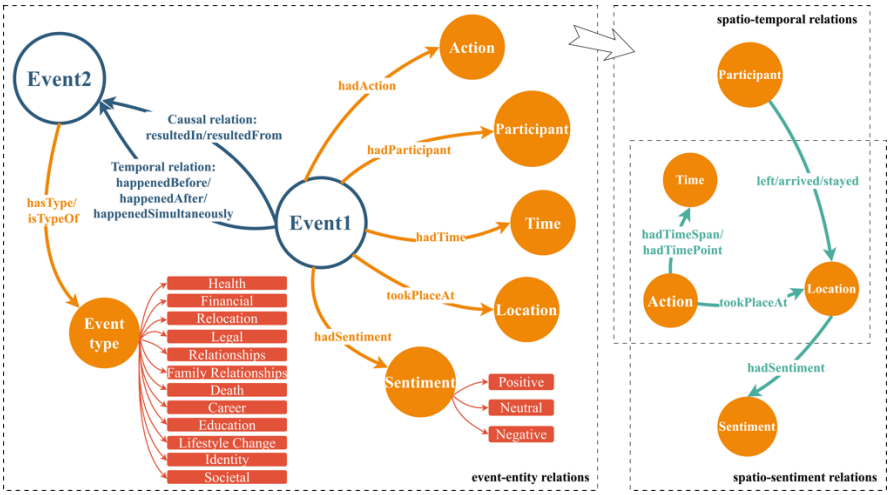


Figure 1. The ontological framework of diaspora oral archives

Event evolutionary graph

Since the EEG is event-centered, including two types of nodes (events and entities) and three types of directed edges (between events, between events and entities, and between entities and entities). The structure of the EEG is formally expressed as G, as follows:

$$G = \{(s, r, o) \mid \{s, o\} \in N, r \in P, N = N_{evt} \cup N_{ent}, P = P_{evt-evt} \cup P_{evt-ent} \cup P_{ent-ent}\}$$

Here the nodes in the EEG are represented as N, mainly including event nodes N_{evt} and entity nodes N_{ent}. Meanwhile, the edges in the EEG are represented as P, mainly including relations $P_{evt-evt}$ between events, relations $P_{evt-ent}$ between events and entities, and relations $P_{ent-ent}$ between entities. In the EEG, event nodes, entities nodes, and their relations are determined by the knowledge representation model.

(1) Events.

An event refers to an objective event or state change consisting of one or more action characteristics participated by one or more arguments in a specific time period or a specific region. Considering the ontological structure, the event is formally expressed as follows:

$$e = (A, P, T, L, S, ET)$$

Here, 'e', 'A', 'P', 'T', 'L', 'S', 'ET' represents Event, Action, Participant, Time, Location, Sentiment and Event Type.

(2) Event relations.

The causal and temporal relations are the most common relations between events (Caselli and Vossen, 2017). Temporal relations can be defined as before, after, simultaneous, begins, ends, etc. The causal relation can be described in precise semantics or hidden in logic, which means it could be inferred (Liu et al., 2021).

Event extraction workflow

We construct an event extraction workflow to construct the information as described in the knowledge representation model. There are four main components for the EE workflow:

(1) Key events and entities extraction.

The real-world events have different granularities, from the top-level themes to key events and then to event mentions corresponding to concrete actions (Zhang et al., 2022). In this context, we propose a new task, key event detection at the intermediate level, which aims to detect event blocks. Each archive is divided into several event blocks manually according to contents. Considering the spatial role that places play and the study's objectives, entities labeled as 'GPE' (Geo-Political Entity) are extracted from each event block. Events associated with placenames are identified as key events. Additionally, other entities related to specific locations, including Participant, Action, and Time, are also extracted.

(2) Event classification.

The event type classification task is based on a taxonomy and automatic classifier. The Major Life Events Taxonomy is employed as the taxonomy of diaspora key events in this research. It is a U.S.-based list of major life changes that people experience (Haimson et al., 2021). In this taxonomy, '*life events*' are used as an umbrella term to encompass life experiences involving both moments and processes of change. It can be used to classify diasporas' key events in their lives into 12 categories: Health, Financial, Relocation, Legal, Relationships, Family Relationships, Death, Career, Education, Lifestyle Change, Identity, and Societal.

(3) Sentiment analysis.

The connection between sentiment and space is particularly significant, as certain locations often evoke distinct emotional responses. Sentiment analysis concerning places can reveal their intuitive feelings about their immigration activities and events. Thus, Aspect-based sentiment analysis (ABSA) is applied to analyze the sentiments ('*positive*', '*negative*' and '*neutral*') of diasporas with the aspects of places (Syamala & Nalini, 2019).

(4) Event relation extraction.

The purpose of event relation extraction is to extract relative clauses and relationship indicator words in the corpus based on syntactic relationships and matching rules, which are tuples in the form of <event A, relation, event B>. This article chooses the commonly used pattern-matching method to extract event relations. It summarizes event relation words and sentences with explicit conjunction (Table 1) and uses a predefined rule base to perform semantic relationship matching.

Relations	Properties	Syntactic pattern	Conjunction
Casual relation	resultedIn/ resultedFrom	{Event_effect} <Conj> {Event_cause}	because, because of, since, as, for, etc.
		{Event_cause} <Conj> {Event_effect}	because of this, so, thus, therefore, consequently, etc.
		<Conj> {Event_effect} {Event_cause}	another important factor / reason of
		<Conj> {Event_cause} {Event_effect}	because, because of, since, as, for, etc.
Temporal relation	happenedBefore/ happenedAfter/ happenedSimulta-neously	{Event_previous} <Conj> {Event_latter}	next, then, after that, later, etc.
		{Event_latter} <Conj> {Event_previous}	previously, prior to this
		{Event_simul1} <Conj> {Event_simul2}	simultaneously
		<Conj>{Event_latter} {Event_previous}	before
		<Conj>{Event_previous} {Event_latter}	after

Table 1. The syntactic pattern and conjunction of event relation pattern

Analysis and results

Data source

The author analyses 42 English transcripts of The South Asian Oral History Project (SAOHP) at the University of Washington Libraries. The SAOHP represents one of the first attempts in the U.S. to record pan-South Asian immigrant experiences in the Pacific Northwest using the medium of oral history. The SAOHP is marked by key historical events that drew South Asians to the United States. These interviews include important events in diasporas' lives reflecting religious, linguistic, occupational, and gender diversity and provide rich insight into the changing experiences of South Asians in the Pacific Northwest. The interviews are in English digitalized and transformed into textual transcripts. Each interview lasts more than 60 minutes and the transcript contains over 20 pages of text. These transcripts were formatted as CSVs separated into chunks of text.

Model evaluation

The spaCy library, DeBERTa and GPT3.5 are selected to conduct the event extraction workflow: entity extraction, event type extraction, sentiment analysis, event relation extraction. Precisely, the spaCy library, and GPT3.5 are conducted on entity extraction and event relation extraction. While in terms of the limitation of the spaCy library, only GPT3.5 are applied to event type extraction and sentiment analysis. Besides, a DeBERTa-based ABSA classifier is also utilized as a supervised machine learning method in sentiment analysis. About 900 manually annotated

samples, are used as the baseline for the evaluation. The performance of these models is assessed using standard NLP metrics such as precision, recall and F1 score. The evaluation results are shown in Table 2.

Tasks	Methods	Evaluation metrics		
		Precision	Recall	F1
Manual annotation		1	1	1
Entity extraction	The spaCy library	42.37%	52.49%	46.89%
	GPT3.5	74.26%	86.80%	80.04%
Event type extraction	GPT3.5	69.75%	77.49%	73.42%
Sentiment analysis	GPT3.5	89.38%	92.39%	90.86%
	DeBERTa	93.54%	92.39%	92.96%
Event relation extraction	The spaCy library	30.25%	35.25%	32.56%
	GPT3.5	12.42%	10.49%	11.37%

Table 2. The evaluation of chosen models

The results demonstrate GPT-3.5 achieves the highest performance in the entity extraction task, while it also demonstrates good performance in event type extraction. For sentiment analysis, the DeBERTa-based ABSA model delivers the best results. However, for event relation extraction, manual validation and complementary methods remain necessary to ensure accuracy.

For each task, the model with the best performance will be utilized to complete all extraction and classification processes. And the extracted data is checked and supplemented through manual inspection. There are 1533 key events determined altogether. And a total of 1043 temporal relationships and 632 causal relationships are ultimately identified.

SEEGraph construction

The proposed event extraction workflow is applied to extracted defined entities, events, and relations, which are imported into the Neo4j database to realize the mapping of knowledge framework and ontology classes to properties and instances. To sum up, there are 1533 event nodes N_{evt}, 19435 entity nodes N_{ent}, 1675 relations $P_{evt-evt}$ between events, 25920 relations $P_{evt-ent}$ between events and entities, and 14825 relations $P_{ent-ent}$ between entities. Hence, the SEEGraph is generated which provides a structured framework for representing and reasoning about diasporas' events.

Event evolution analysis

The SEEGraph traces the sequences, patterns, and trajectories of events influencing diaspora movements, transformations, and experiences over time. For example, Seattle had its large share of Indians who were predominately engineers who worked for Boeing in the 1960s and 1970s. The job offers from Boeing have become a key facilitator for diasporas to immigrate during that period. As for AM (Figure 2), after getting a job at Boeing, he moved to Seattle. He made friends with colleagues who are also Indian diasporas and was admitted to the engineering college which collaborated with Boeing company.

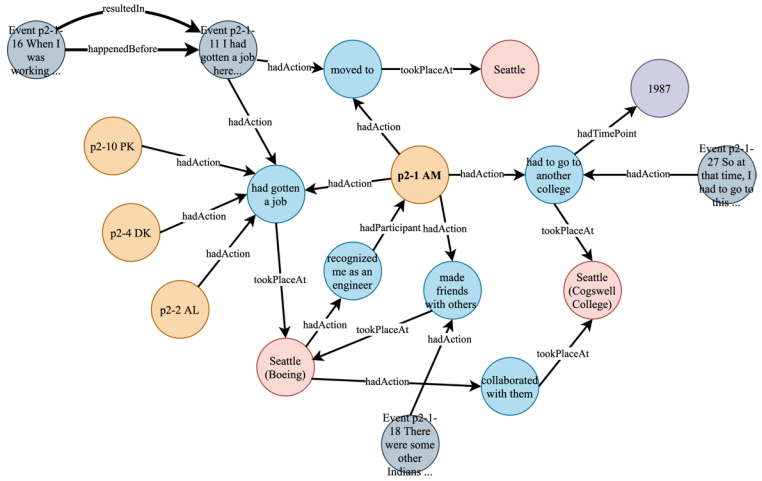


Figure 2. The event evolutionary process according to job offers from Boeing

Spatio-temporal analysis

The spatio-temporal analysis creates a multidimensional representation of spatial movements and temporal changes. The whole picture of interviewees' immigration activities is revealed in Figure 3. The nodes are the places they have stayed for more than one year and the connecting curves show the flows and years of immigration. We can conclude four types of immigration modes: single-directional, round trip, multi-directional, and internal trip.

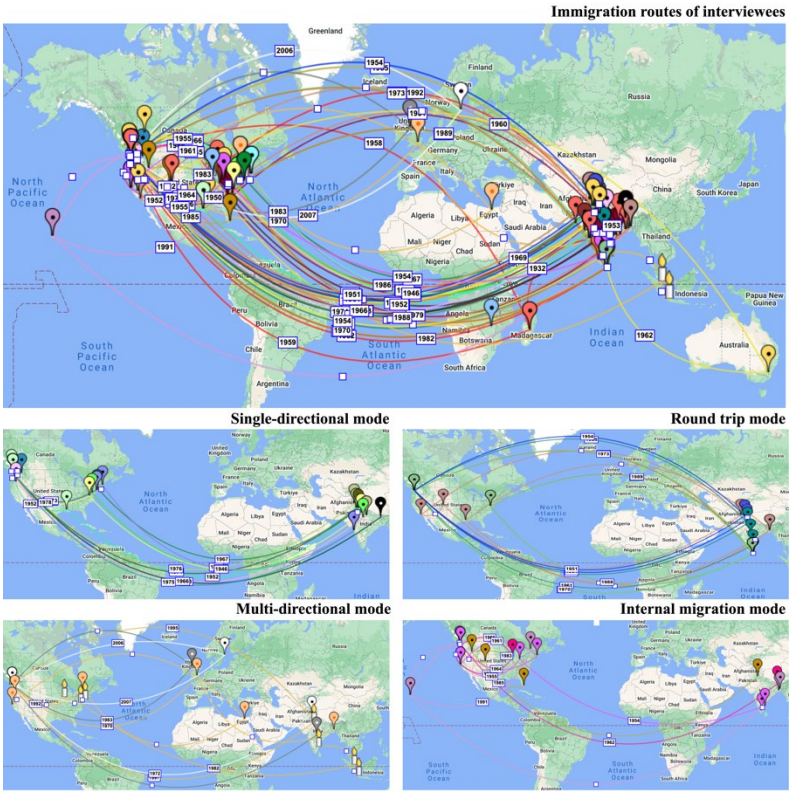


Figure 3. Interviewees' immigration routes and modes

Spatio-sentiment analysis

The spatio-sentiment analysis creates a rich and context-aware representation of the diasporas' sentiment towards geographical entities, especially their destination and homeland. As is shown

in Figure 4, In terms of their sentiment toward the most frequently mentioned locations, their comments on the USA and American cities are one-sided, and positive perspectives are much more than negative ones. According to their sentiment towards their homeland, the feelings are more complicated and conflicted. An almost equal quantity of positive and negative opinions indicates a mixed reviews attitude.

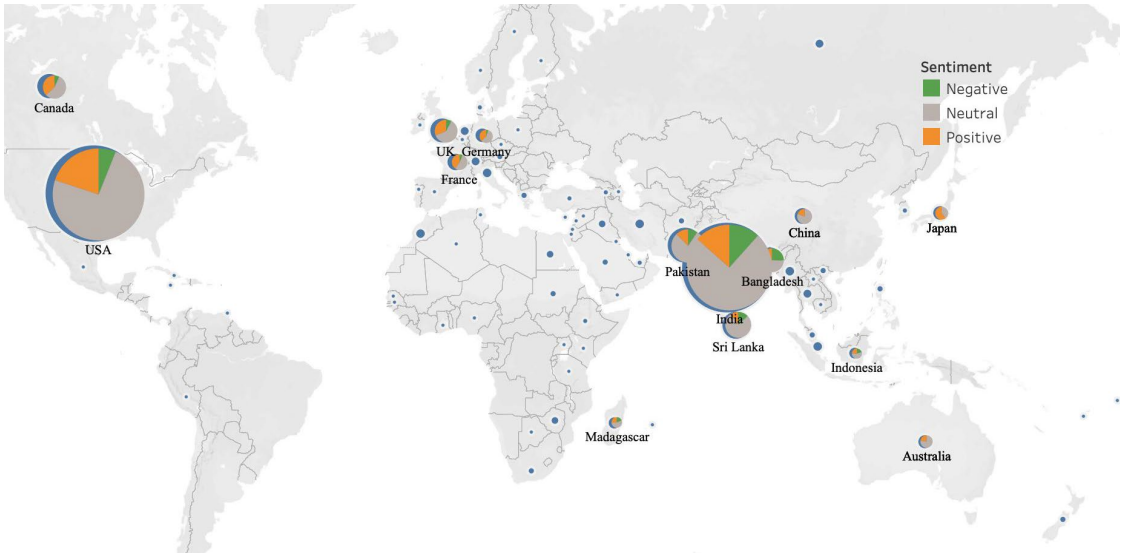


Figure 4. The diasporas' sentiment toward main locations

Discussion and conclusion

The analysis of diaspora oral archives serves as a pathway to represent diversified voices in repositories. In this research, the ensemble framework is introduced not only to realize the semantic, correlation, and structured expression of knowledge units such as events, entities, and attributes but also to clearly reveal the dynamic evolution rules and patterns between historical and personal events. this study has innovations and contributions in the following aspects:

Firstly, unlike previous event extraction conducted using relatively high-quality data such as media articles (Li et al., 2018; Rudnik et al., 2019), the presented study processes the information extraction of diaspora oral archives transformed into structured text. Specifically, the dataset used in this study is differentiated in the following aspects: (1) longer and complicated context in the interview conversation with imbalanced length, (2) complex event content and relations due to interviewees' different narratives, and (3) including many idiomatic expressions, slang, colloquialisms, or repeated expressions in oral narratives. To solve this, the ontological model is proposed to normalize the domain of diasporas' major life events. Instead of defining event as event trigger words and event arguments as in previous research (Guan et al., 2022), this study refines it as action, participant, time, location, sentiment, and event type. So, the extracted information in the event is more abundant and comprehensive. It fills research gaps in event detection which mainly focuses on political events or news corpora. Besides, it's the first time that the EEG combined with NLP pipelines to analyze diaspora oral archives. The performance of spaCy, LLMs, etc. has been examined in each task. The ensemble framework with multiple models is expected to be used in the processing of other diaspora oral archives and low-quality textual transcripts of oral materials in the future.

Secondly, the sentiment is involved, which is also first introduced into the EEG and diaspora oral archives. This study helps to reveal the reasons and results of the diasporas, how they are influenced and exert an influence on the mobility of the migration, and how they build up a sense of belonging toward the destination country. This led to intense confrontations, with both positive

and negative emotions. Though in this research the sentiment is coarse-grained, it can be seen as the first attempt to understand and respond appropriately to diasporas' emotional reactions. With embedded sentiment, diasporas' affective assessment of places and events deepens the understanding of their immigration activities.

Acknowledgements

I would like to express my gratitude to Prof. Charles Jeurgens for his careful reading of the manuscript and his valuable feedback and suggestions.

The oral history transcripts used in this study are part of the South Asian Oral History Project (SAOHP) collection at the University of Washington Library. While these materials are publicly accessible, their copyright remains with the creators (interviewees and interviewers). This study uses the data strictly for academic research purposes in compliance with the principles of '*Fair Use*.' The data has been anonymized, and no identifiable personal information is disclosed. Any results or findings are derived from analytical methods and do not reproduce or distribute the original content of the transcripts.

About the author

Jing Zhou is doctoral candidate in the School of Information Management, Wuhan University. Her research interest focuses on digital humanities, knowledge organization and natural language processing. She can be contacted at zhoujingwinky@whu.edu.cn.

References

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023. https://doi.org/10.48550/arXiv.2302.04023.

Bekiari, C., Bruseker, G., Doerr, M., Ore, C. E., Stead, S., & Velios, A. (2021, April). Volume A: Definition of the CIDOC conceptual reference model. Version 7.1. (https://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.1.2.pdf).

Brinkhurst, E. (2012). Archives and Access: Reaching Out to the Somali Community of London's King's Cross. Ethnomusicology Forum, 21(2), 243–258. https://doi.org/10.1080/17411912.2012.689470

Brown, M. & Shackel, P. (2023). Text Mining Oral Histories in Historical Archaeology", International Journal of Historical Archaeology, 27, 865–881. https://doi.org/ 10.1007/s10761-022-00680-5.

Caswell, M. & Mallick, S. (2014). Collecting the easily missed stories: digital participatory microhistory and the South Asian American Digital Archive. Archives and Manuscripts, Taylor & Francis, 42(1), 73–86. https://doi.org/10.1080/01576895.2014.880931.

Caselli, T. & Vossen, P. (2017, August). The event storyline corpus: A new benchmark for causal and temporal relation extraction. In Proceedings of the Events and Stories in the News Workshop (pp. 77-86). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2711.

Dominguès, C., Jolivet, L., Brando, C. & Cargill, M. (2019). Place and Sentiment-based Life story Analysis. From the Spanish Republican Army to the French Resistance. Revue Française Des Sciences de l'information et de La Communication, 17. https://doi.org/10.4000/rfsic.7228.

Gottschalk, S., Kacupaj, E., Abdollahi, S., Alves, D., Amaral, G., Koutsiana, E., ... & Thakkar, G. (2023). OEKG: The open event knowledge graph. arXiv preprint arXiv:2302.14688. https://doi.org/10.48550/arXiv.2302.14688.

Grant, K.A. (2020). Affective Collections: Exploring Care Practices in Digital Community Heritage Projects (Unpublished master thesis). University of Alberta, Alberta, Canada

Guan, S., Cheng, X., Bai, L., Zhang, F., Li, Z., Zeng, Y., ... & Guo, J. (2022). What is event knowledge graph: A survey. IEEE Transactions on Knowledge and Data Engineering, 35(7), 7569-7589. https://doi.org/10.1109/TKDE.2022.3180362.

Haimson, O.L., Carter, A.J., Corvite, S., Wheeler, B., Wang, L., Liu, T. & Lige, A. (2021). The major life events taxonomy: Social readjustment, social media information sharing, and online network separation during times of life transition. Journal of the Association for Information Science and Technology, 72(7), 933–947. https://doi.org/10.1002/asi.24455.

Jones, M. (2019). Archiving the trauma diaspora: Affective artifacts in the higher education arts classroom. Marilyn Zurmuehlen Work. Pap. Art Educ, 2019, 1-14.

Li, Z., Ding, X., & Liu, T. (2018, July). Constructing narrative event evolutionary graph for script event prediction. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden (pp. 4201-4207). AAAI Press.

Liu, Y., Tian, J., Zhang, L., Feng, Y., & Fang, H. (2021). A Survey on Event Relation Identification. In Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence: 5th China Conference, CCKS 2020, Nanchang, China, November 12–15, 2020, Revised Selected Papers (pp. 173-184). Springer Singapore. https://doi.org/10.1007/978-981-16-1964-9_14.

Ozdemir, C. & Bergler, S. (2015). CLaC-SentiPipe: SemEval2015 subtasks 10 B, E, and task 11. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (https://aclanthology.org/S15-2081.pdf).

Pessanha, F., & Salah, A. A. (2021). A computational look at oral history archives. ACM Journal on Computing and Cultural Heritage (JOCCH), 15(1), 1-16. https://doi.org/10.1145/3477605.

Rieping, H.A. (2022). Audio Segmenting and Natural Language Processing in Oral History Archiving. (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Massachusetts, U.S.A.

Roeschley, A. & Kim, J. (2019). “Something that feels like a community”: the role of personal stories in building community-based participatory archives. Archival Science, 19, 27–49. https://doi.org/10.1007/s10502-019-09302-2.

Rudnik, C., Ehrhart, T., Ferret, O., Teyssou, D., Troncy, R., & Tannier, X. (2019, May). Searching news articles using an event knowledge graph leveraged by wikidata. In Companion proceedings of the 2019 world wide web conference (pp. 1232-1239). Association for Computing Machinery.

Syamala, M., & Nalini, N. J. (2019). A deep analysis on aspect-based sentiment text classification approaches. International Journal of Advanced Trends in Computer Science and Engineering, 8(5), 1795-1801.

Thomson, A. (1998). Fifty years on: An international perspective on oral history. Journal of American History, 85(2), 581. https://doi.org/10.2307/2567753.

Yang, H., Zeng, B., Xu, M. & Wang, T. (2021). Back to Reality: Leveraging Pattern-driven Modeling to Enable Affordable Sentiment Dependency Learning. ArXiv Preprint. (https://www.researchgate.net/profile/Heng-Yang-

17/publication/355391949_Back_to_Reality_Leveraging_Pattern-driven_Modeling_to_Enable_Affordable_Sentiment_Dependency_Learning/links/6189682107be5f31b7590ae3/Back-to-Reality-Leveraging-Pattern-driven-Modeling-to-Enable-Affordable-Sentiment-Dependency-Learning.pdf).

Zhang, Y., Guo, F., Shen, J., & Han, J. (2022, August). Unsupervised key event detection from massive text corpora. In Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining (pp. 2535-2544). Association for Computing Machinery. https://doi.org/10.1145/3534678.3539395.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Innovative practice of archival data development workflow in the AGI era: a case study of scientist archives project

Yaming Fu, Jie Song, Xinran Zhang, and Jingyun Bi

DOI: https://doi.org/10.47989/ir30iConf47335

Abstract

Introduction. The advent of large language models (LLMs) presents a transformative opportunity for the field of archival science, offering advanced capabilities in intelligent information processing, semantic search, and more. These innovations address critical challenges posed by the exponential growth in archival materials and the increasing demand for efficient data analysis. Traditional archival workflows, often rely on manual description and optical character recognition (OCR), struggle with the complexities of unstructured digital data, especially in the context of digitized historical archives and manuscripts.

Method. This paper explores a novel archival workflow through the case study of the scientist archives project, integrating human-machine collaboration and leveraging technologies such as open-sourced archival databases, IIIF-supported OCR environments, and advanced LLMs, including classic retrieval-augmented generation (Classic RAG) and graph-based retrieval-augmented generation (GraphRAG).

Analysis. A Scientist Archives project is analysed, which utilises AGI era technologies to mine and manage the archives.

Results. By embracing these technologies, the proposed approach seeks to revolutionize archival management, enhancing both efficiency and the depth of content revelation in the AGI era.

Conclusion(s). This study contributes to the ongoing discourse on the intelligent transformation of archival practices, providing a roadmap for future archival data mining and management.

Introduction

Since 2018, the archival community has actively discussed the impact of artificial intelligence and digital technologies, and research on the digitalization and intelligent transformation of archival management has attracted much attention (Rolan, 2021). By late 2022, the release of ChatGPT marked a significant milestone in the development of large language models (LLMs), fueling growing discussions and heightened expectations for the era of artificial general intelligence (AGI) (Emmert-Streib, 2024). As a core and controversial concept in the field of computer research, AGI refers to the AI systems that exhibit comparable intelligence to human intelligence in a wide range of tasks and situations, and such systems usually have general task-solving capabilities and continuous self-learning capabilities and '*exhibit brain-like intelligence characteristics in cognition and decision-making*' (Morris, 2023). LLM shows versatility in natural language processing, provides a window for the public to approach AGI, and serves as an important milestone on the road to AGI. As a generative natural language processing model for large-scale corpus training, '*LLM has the functions of automatic text generation, intelligent information processing, semantic search and recognition, intelligent image generation, etc.*' (Naveed, 2023), which provides a promising research direction for archival data mining and archival management, and triggers deep thinking on a more '*intelligent*' transformation of archival work.

The surge in archival materials, the growing demand for efficient methods of analysing archival data and in-depth archival mining have brought unprecedented challenges to traditional archival workflow. Manual description processes are time-consuming, labour-intensive, and prone to errors, particularly when dealing with large volumes of archival materials. Optical character recognition (OCR) software is widely used to extract text from digital images, facilitating the creation of searchable text files (Bukhari, 2017). However, digital archival data is often unstructured, complicating retrieval and utilization. This is particularly true for digitized historical archives and manuscripts, which are difficult to recognize, analyse, and describe. Moreover, traditional workflows fall short in granular content revelation, failing to meet users' diverse information needs in the new era.

This project uses the archival materials of Chen Taosheng (In Chinese: 陈駒声), the founder and pioneer of modern industrial microbiology in China, as a case study to explore a novel archival workflow that integrates human-machine collaborative models in managing scientist archives. The proposed workflow leverages advanced technologies, including an open-sourced archival database, the international image interoperability framework (IIIF) supported work environment, and LLMs, particularly classic retrieval-augmented generation (Classic RAG) and graph-based retrieval-augmented generation (GraphRAG). The project aims to revolutionize archival workflows, enhancing the overall efficiency and quality of archival management in the era of AGI.

Problem statement

Traditional archival workflows are insufficient to meet the demands of modern archives management. This is not only due to the unreliable manual processing and cataloguing but also the fragmented use of independent tools for different archival tasks. These tools include but are not limited to '*metadata management tool, OCR tool, ontology design tool and knowledge graph tool*' (Koch, 2019; Philips, 2020). In addition, traditional workflows lack the granularity needed to reveal the full depth of archival content, but in the meantime, the authenticity and security of the archival data remain, which is key to archival management. Therefore, the research problems of this project are:

First, design and implement an IIIF-based OCR environment to support archival transcription, recognition, comparison, and description. IIIF, as '*a set of open standards for delivering high-quality, attributed digital objects online at scale*' (Link: https://iiif.io), could provide standardized image

processing protocols, enhancing image-formatted archival data management, improving the efficiency, and ensuring accuracy of archival data.

Second, formulate a workflow for managing scientists' archives, organize scientists' archives, photos, literary works, monographs, patents, and relevant documents, index them in the open-sourced archival workspace ArchivesSpace, and integrate them with LLM tools to facilitate efficient access and utilization.

Third, build Classic RAG and GraphRAG to deeply mine and analyse archival data, improve the interpretation ability of archival data with scanned archival images, and design the 'chat' function with AIGC, forming a new paradigm of archival utilization.

With the above three research goals, the project aims to develop an AGI-Era Archival workflow solution, an efficient and reliable framework to promote the intelligent processing of archival description, analysis, development, and utilization.

Methodology

Research process

This work originated in early 2023 with the call to collect, organize and preserve the archival materials of scientists, especially the ones that were kept by individuals and the ones that were in bad condition. The aim was to collect, digitise, analyse, and make use of the scattered and deteriorating archival materials related to the lives and work of scientists, as well as to excavate the values hidden in the archives with available technologies. In October 2023, a project member interviewed a close friend of Chen Taosheng, a founder and pioneer of modern industrial microbiology in China and a key figure in establishing the department of bioengineering at the predecessor of Shanghai University. During this interview, the project member obtained valuable family archival materials. Subsequently, project members visited the Shanghai municipal archives, and the Shanghai University archives, acquiring more materials related to Chen Taosheng. With the number and types of archival materials growing, an urgent need has emerged to design an efficient workflow to manage, describe, interpret, and analyse those archival materials safely without damaging the original files' quality and meaning. Thus, the project takes Chen Taosheng's archival materials as a case study to explore the development of an archival workflow for archival data in the AGI era.

An innovative archival development workflow in the AGI era should enhance efficiency while maintaining data authenticity, accuracy, and security. This project, an AI-integrated working environment based on the IIIF framework and embedded with OCR tools, was designed to facilitate processing image-based archival data. It also utilized LLM technology to assist in analysing massive archival raw images and their corresponding texts. Classic RAG and GraphRAG were built to realize the '*chat*' function with the scientist and form scientists' knowledge graphs, scientists' footprint maps, and scientists' chronology with the help of artificial intelligence, thereby preserving and extending their research contributions. Through this integrated platform, tasks such as archival recognition, classification, transcription, description, comparison, and analysis were completed in a unified environment, significantly improving the efficiency of archival development while ensuring data accuracy.

Data and ethical statement

All archival materials utilized in this project were obtained with permission from four sources, including:

(1) the close friend of Chen Taosheng: The materials are mainly published magazines and newspapers related to the scientist, essays published or mentioned by Chen Taosheng himself, life photos of Chen Taosheng, and his unpublished poetry collection '*Guanwei collection*' (in Chinese: 观微集).

(2) Shanghai municipal archives: the materials contain his views on scientific research, his career and work proposal and its reply, his work introduction letters, and materials for selecting advanced individuals in the science field.
(3) Archives of Shanghai University: the materials include the forms filled in by him during his work at the predecessor of Shanghai University, his thoughts and opinions, the relevant school records, and his relevant letter of appointment.
(4) Open data from Shanghai Library: records of other celebrities and scientists related to Chen Taosheng; records of the institutions and places that relate to him.

The types of archival data include original archival files, old photographs, manuscripts, newspaper clippings, journal articles clippings. These materials were acquired with explicit agreements permitting their digitization and use only for academic purposes.
The subsequent data processing and experiments were conducted within the university laboratory under the oversight and management of the university's office of scientific research management, safeguarding the integrity and authenticity of the archival materials throughout the project.

Research process and methods

In the scientist archival data development workflow project, the main research path is '*physical form of archives—digital archival images—textual archival data—archival knowledge*' (see Fig. 1).

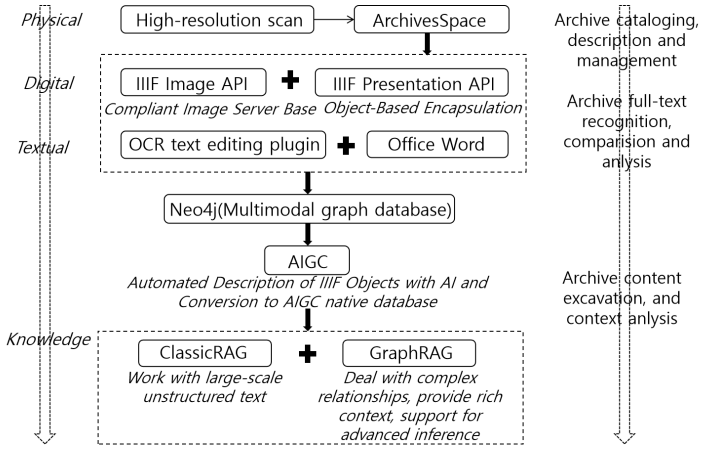


Figure 1. Research process and methods

The workflow began with digitizing physical archives and records, utilizing high-resolution scanners to convert archival documents, photographs, and other archival materials into digital formats. These digitalized files were then organized and catalogued using the open-sourced archival management system, ArchivesSpace, where structured metadata and standard archival descriptions to enhance data retrieval and operability (see Figure 2). During this process, raw images were integrated through IIIF objects and recorded as digital object instances corresponding to physical items.

'*ArchivesSpace is an open-source, browser-based archival information management software application*' (Link: https://archivesspace.org). The underlying descriptive standard it uses is DACS (Describing Archives: A Content Standard). ArchivesSpace allowed the users to add, arrange, describe, evaluate materials, manage locations, control permissions, and link subject tags when storing archive metadata. DACS is a foundational statement of content standards and principles for archival organization and description within the United States archival community. It includes descriptions of archival material and authoritative records representing the people and organizations that created them (SAA, 2022). DACS contains 25 archival elements, many encoded with multiple EAD tags.

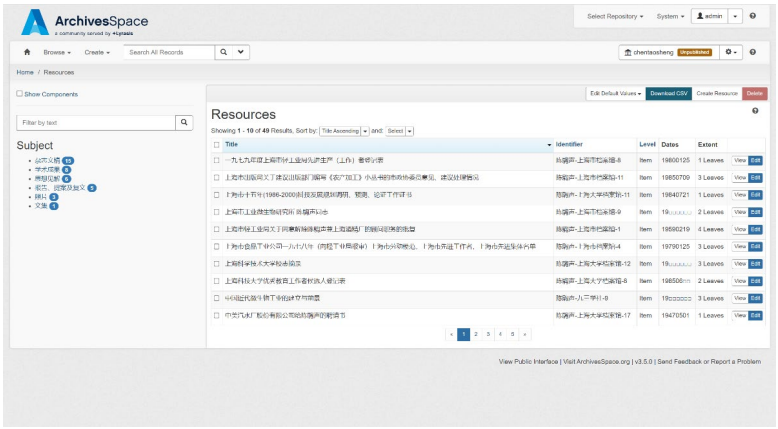


Figure 2. ArchivesSpace archival management system

Then IIIF was integrated with the OCR text editing plugin in the Office Word environment (see Figure 3) to realize a commonly used archival work environment and the seamless linking of raw images with their corresponding OCR texts, ensuring comprehensive documentation and accessibility.

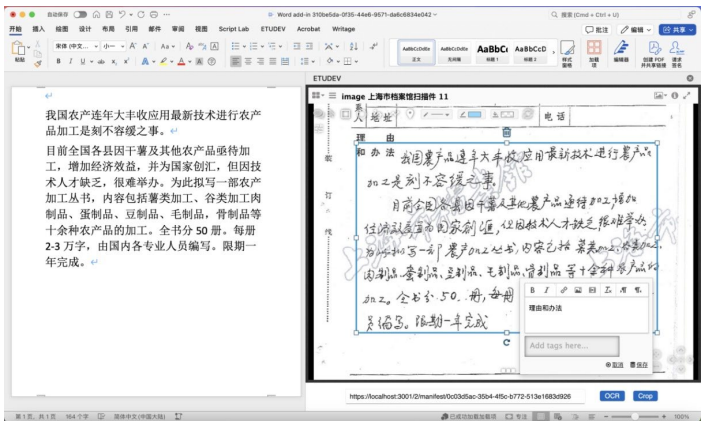


Figure 3. IIIF integrated word environment with OCR plugin

Archival images (unstructured data) are transformed into structured and semi-structured data through the IIIF-based OCR tool. Structured data refers to tabular data that can be directly stored in a structured database, while semi-structured data refers to data in a loose key value format such as JSONL format. Traditional manual description tasks are thus converted into data mapping tasks that can be automatically completed by LLMs. Once a high-quality electronic archival database is constructed, it can serve as a private AIGC database for Classic RAG or GraphRAG needed by LLMs, targeting at content mining, content analysis, and other tasks that can reveal the value of archives.

Findings

Large archival digital images solutions

Using scanned archival images as archival data posed several technical challenges, including handling large quantities of images, managing complex image-text interactions, ensuring description accuracy, and maintaining archival security when employing LLMs. To address these challenges, the project adopted the IIIF framework as a standard, breaking down image development into three independent yet progressive workflows.

First, the project implemented an IIIF Image API as a compliant image server base (see Figure 4 and Figure 5). The Image API defined pixel-level operations for images, forming the foundation for dynamically presenting archival images, particularly high-definition images. It also enabled real-time acquisition and manipulation of finely indexed images. The project incorporated a series of engineering technologies, including container orchestration and serverless computing, to ensure system performance meets business needs when handling large quantities of high-definition images.



Figure 4. IIIF Image API

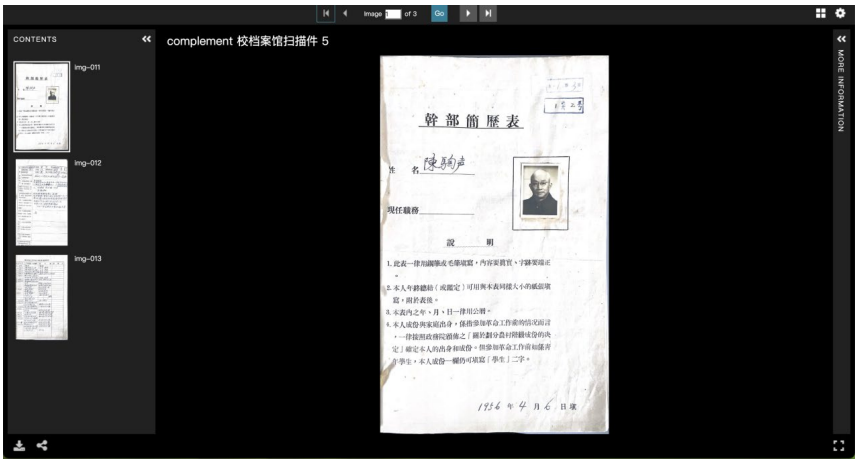


Figure 5. An example of IIIF Image API

Second, the project utilized the IIIF presentation API for object-based encapsulation and OCR Text integration. The project leveraged the flexible definition methods of the presentation API to combine individual images into logically related image objects and described metadata in multiple languages at different levels (see Figure 6). Given the diverse layout of images, the project incorporated the ability to select appropriate areas freely and to apply different OCR models for handwritten and printed texts. The project developed an OCR text editing plugin that combines IIIF and office word, enhancing annotation flexibility and solving the problem of standardized storage of annotation results, enabling quick associations between texts and images during retrieval.

form	collection	manifest	item	range	annotation
Manuscript	Filing	File	Page	Catalog index	OCR full—text
Book	Filing	File	Page	Catalog index	OCR full—text
File sheet		Copy	Page	Spread index	OCR sheet
Photo		Collection	Piece		OCR local annotation

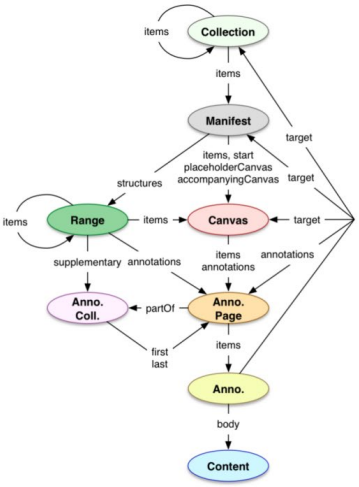


Figure 6. Flexible definition methods of the presentation API

This solution preserves the originality of archival materials, respects the existing form of archival materials, gives them the possibility of being used as '*data*' and '*evidence*', and ensures that archival materials can be preserved for a long time.

Archival data AIGC knowledge base and LLM solutions

By leveraging LLMs' language reconstruction and mining capabilities, the project significantly enhanced the explanatory power of archival data, establishing a new paradigm for archival utilization within the new technical environment. '*RAG (retrieval enhanced generation) technology is evolving, including the Classic RAG and Graph RAG, representing a significant advance in information retrieval and knowledge representation*' (Omrani,2024). This project explored the two approaches and built a chat function allowing users to 'chat' with the scientist (see Figure 9).

The Classic RAG system (see Figure 7) first converted documents into Markdown format through OCR. Using the hierarchical structure of this format, the system employed an overlapping sliding window method to segment text while maintaining semantic integrity. Each segmented text block was then converted into a JSONL object containing metadata such as text content and URLs. These objects are then converted into embedding vectors stored in a vector database. During retrieval, the system augmented the context window by locating similar text in the database through nearest neighbour search.

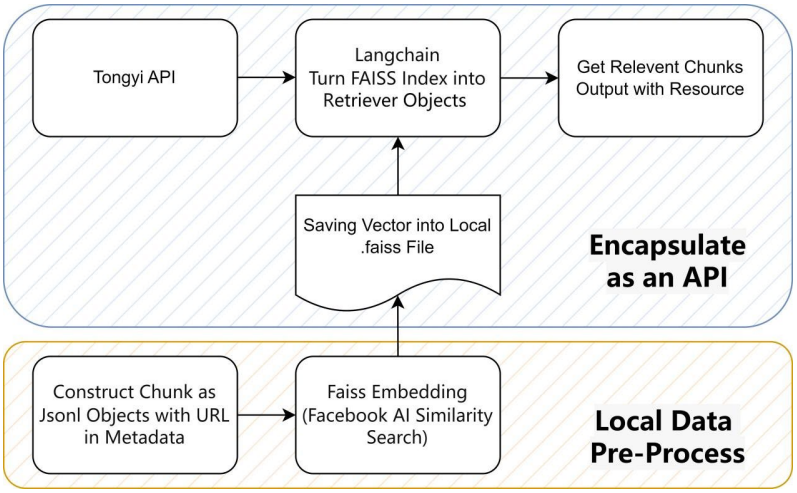


Figure 7. Classic RAG

GraphRAG followed this process but introduced an additional layer of inference. The LLM first performed inference operations on each sentence in the dataset, extracting entities and relationships. Named entity recognition (NER) was used to identify entities within the text and determine the relationships between entities, as well as the strength of those relationships. Then, the extracted relationships were formed into a knowledge graph, which consisted of nodes (entities) and edges (relationships) (see Figure 8).

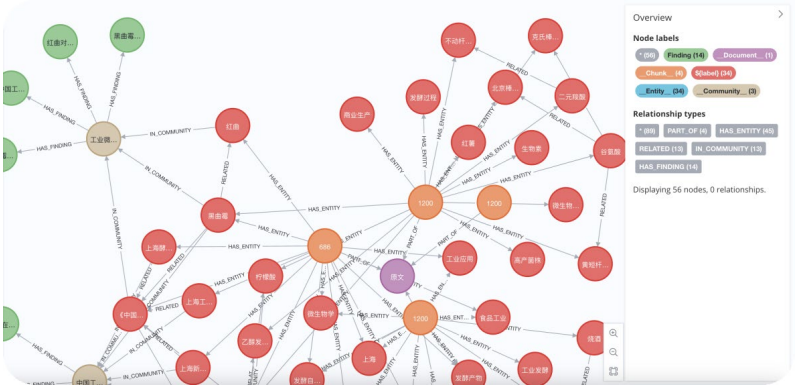


Figure 8. Knowledge graph based on GraphRAG

While the Classic RAG primarily relied on vector databases to process unstructured text, the Graph RAG established a super-large-scale vocabulary by introducing knowledge graphs that equated entities and relationships with words. This enabled a more accurate understanding of content and more accurate search results.

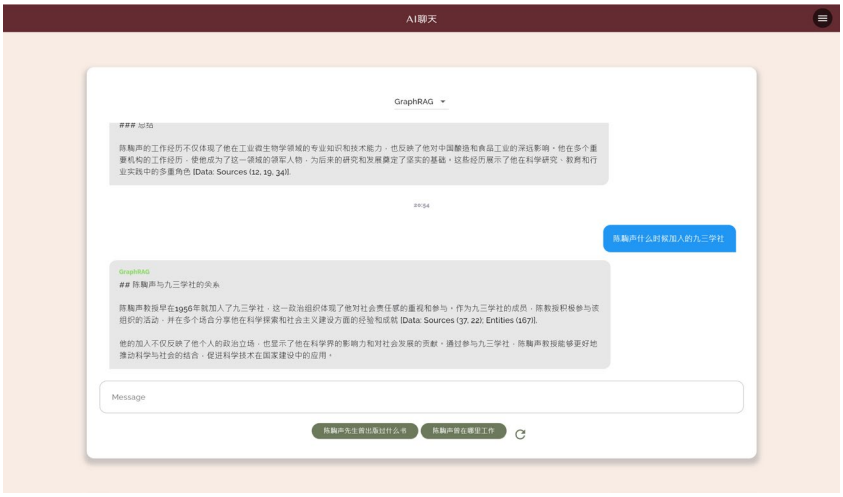


Figure 9. Chat interface of the project

Towards a human-machine archival workflow

The Chen Taosheng archival data development project successfully implemented a systematic workflow by utilizing the open-source archival management system ArchivesSpace, along with an IIIF integrated word environment with OCR plugin, for comprehensive archival data cataloguing, describing, recognition, comparison, and management. Additionally, by extending the LLM with Classic RAG and GraphRAG, the project successfully enabled AIGC-generated texts to refer to original images. This technique not only effectively addressed potential hallucination issues in LLMs but also significantly enhanced the credibility of LLM responses by presenting the original images, thereby maximizing the utility and reliability of archival resources.

The integrated IIIF and OCR tool-based work environment ensured the automated generation of full-text data from archival materials and enhanced the efficiency and accuracy of the process through human-machine collaboration. The automation of repetitive tasks reduced the workload on archivists, allowing them to focus on more complex aspects of archival work. This hybrid approach not only improved efficiency but also reduced the risk of human error. The adoption of LLM supported the archival workflow in a good way by enabling the explanation ability of archival materials and enriching users' interaction with archives.

Lessons and future research

The integration of AIGC texts with original images through IIIF objects effectively addressed the hallucination issue in LLMs, significantly enhancing the credibility of LLM responses. However, several challenges were encountered during the project experiment when interacting with the LLM. Firstly, when answering questions requiring large chunks of content, the original large chunks of content were often truncated (see Figure 10 as an example). This truncation hindered the LLM's ability to identify relationships between keywords, leading to responses that, although traceable, were often speculative and lacked credibility. To solve this issue, the size of the text chunks and the length of overlapping segments were typically increased to make the semantic relationships more complete. However, this approach proved insufficient for large-scale overviews unless the corpus was rich enough to provide summative content. This limitation underscored the need for GraphRAG.



Figure 10. Large chunks of content may be truncated in the answering the question.

Secondly, ambiguity in naming entities posed another significant issue. For example, '*Mr. Chen*' and '*Professor Chen*' in the literature both referred to Chen Taosheng, however, these designations were sometimes incorrectly identified as different entities when constructing knowledge graphs. This problem was further exacerbated by OCR errors (see Figure 11 as an example). To solve this problem, cluster-based entity disambiguation is required, i.e., identifying the real-world entity to which ambiguous referents were directed by leveraging the principle that the same referent typically appeared in approximate contexts. Clustering algorithms were then applied to disambiguate these entities.

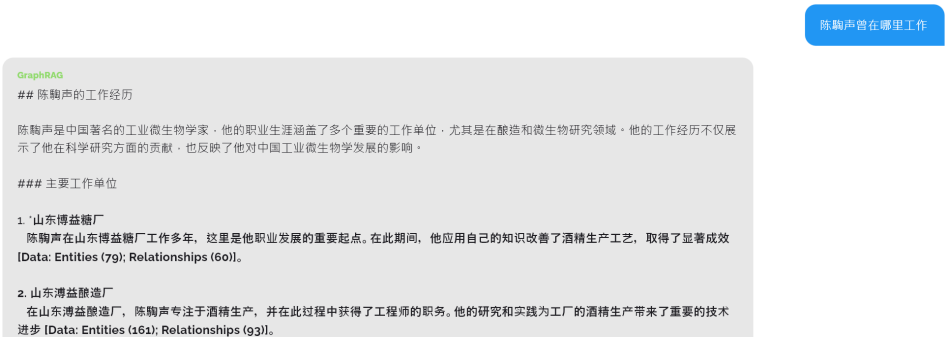


Figure 11. Ambiguity in naming entities.

Future directions for the project include considering the integration of multi-modal large language models (MLLMs) to deeply mine semantic information from raw archival images beyond text, creating innovative modes for multi-modal archival management and development. The project team also plan to continue to optimize the human-machine collaborative archival development models and technological applications. These efforts aim to explore more efficient data processing and archival management solutions, further enhancing the intelligence and processing capabilities of archival work. Such advancements were expected to not only increase the value and utilization efficiency of archival data but also provide new theoretical support and practical guidance for the development of the archival industry.

Conclusion

This project demonstrated the potential of modern technology in the AGI era in archival management through a case study of the scientist Chen Taosheng's archival materials. By leveraging IIIF, OCR and LLMs technologies, and implementing human-machine collaborative models, an innovative workflow for native image archival data has been developed in this project. These advancements significantly enhanced the efficiency and accuracy of archival development and utilization, providing theoretical and practical guidance for the field.

As the AGI era unfolded, the workflow for managing archives and records underwent a transformative evolution. The collaboration between humans and machines proved crucial in

navigating the complexities of the AGI era. Human expertise in archival science ensures the contextual accuracy and ethical handling of records, while machine capabilities significantly expedite the processing and organization of vast amounts of data. This symbiotic relationship enhances the overall workflow, leveraging the strengths of both human intuition and machine precision. Ensuring the accuracy and trust of archive documents requires robust validation mechanisms, where human oversight plays a critical role in verifying the outputs generated by automated systems. By fostering a collaborative environment, archivists are able to release the full potential of AGI technologies, driving innovation in archival practices while maintaining the integrity and authenticity of archival records. This balanced approach not only preserves the historical and cultural significance of archives but also paves the way for more advanced and reliable archival management in the future.

About the authors

Dr. Yaming Fu is a Lecturer at School of Cultural Heritage and Information Management, Shanghai University, Shanghai, China. Her research interests include digital humanities, knowledge organization and archives informatization. E-mail: ymingfu@126.com

Jie Song is the Founder and CEO of Intellijourney Co., Ltd., Shanghai, China. His research interests include IIIF and artificial intelligence. E-mail: joe.song@intellijourney.com

Xinran Zhang is an undergraduate at School of Cultural Heritage and Information Management, Shanghai University, Shanghai, China. Her research interests include basic theory of archival science, digital humanities. E-mail: 1654770712@qq.com

Jingyun Bi is an undergraduate at Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. His research interests include application of generative AI in social data, and radio frequency surrogate models for automatic design of active and passive circuits. E-mail: 2537503782@qq.com

References

Rolan, Gregory, et al. 'More human than human? Artificial intelligence in the archive.' Archives and Manuscripts 47.2 (2019): 179-203.

Emmert-Streib, Frank. 'Is ChatGPT the way toward artificial general intelligence.' Discover Artificial Intelligence 4.1 (2024): 1-8.

Morris, Meredith Ringel, et al. 'Levels of AGI: Operationalizing Progress on the Path to AGI.' arXiv preprint arXiv:2311.02462 (2023).

Naveed, Humza, et al. 'A comprehensive overview of large language models.' arXiv preprint arXiv:2307.06435 (2023).

Bukhari, Syed Saqib, et al. 'anyocr: An open-source ocr system for historical archives.' 2017 14th IAPR international conference on document analysis and recognition (ICDAR). Vol. 1. IEEE, 2017.

Koch, Inês, et al. 'Knowledge graph implementation of archival descriptions through CIDOC-CRM.' International conference on theory and practice of digital libraries. Cham: Springer International Publishing, 2019.

Philips, James P., and Nasseh Tabrizi. 'Historical document processing: historical document processing: a survey of techniques, tools, and trends.' arXiv preprint arXiv:2002.06300 (2020).

Link: https://iiif.io

Link: https://archivesspace.org

SAA. Describing Archives: A Content Standard.Society of American Archivists' Technical Subcommittee, https://saa-ts-dacs.github.io/dacs/06_part_I/02_chapter_01.html (2022).

Omrani, Pouria, et al. 'Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement.' 2024 10th International Conference on Web Research (ICWR). IEEE, 2024.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Data feminism in museum collections: mapping principles to practices

Alexandria Rayburn

DOI: https://doi.org/10.47989/ir30iConf47332

Abstract

Introduction. In library, archive, and museum (LAM) collections, there has been much work towards restorative cataloguing and addressing harms found within collection data. However, this work is still being done in ad-hoc ways. As a result, there is both 1) a lack of a comprehensive summary of this work and 2) a gap in connecting this work to similar research in critical data studies. Parallel to this, the recent publication *data feminism*, offers a framework for critically examining data practices through a lens of power and oppression.

Method. This qualitative systemic review maps the seven principles of *Data Feminism* (Examine Power, Challenge Power, Elevate Emotion and Embodiment, Re-think Binaries and Hierarchies, Embrace Pluralism, Consider Context, Make Labor Visible) to existing work happening in museum collections.

Results. While museum practitioners are adept at understanding the limitations of their data schemas and organizing structures, further work is required to incorporate emotion and embodiment into museum data, as the current model relies on external researchers and community members for this.

Conclusion. This review underscores the importance of interdisciplinary work and speaks to the intensive data labor happening in LAM collections; it also offers actionable insights for advancing feminist data practices within museum collections.

Introduction

Library, archive, and museum (LAM) collections are sites where data-intensive work has been conducted for decades (Strasser, 2012; Sepkoski, 2013; Turner, 2020). This data often has problematic legacies rooted in colonialism, imperialism, and patriarchy, making their data equally filled with intersecting biases (Lowry 2023; Philips 2011; Reilly 2018; Turner 2020), yet touted as objective fact. These biases result from the often-obscure methods to which the collections came to be, and through the data classification schemas that reinforce white/patriarchal ideologies while erasing other forms of knowledge in the process (Patin et al. 2020). Many practitioners are invested in repairing these harms in their collection data.

Reparative or critical data work is not unique to museums, however. The recent publication *data feminism* offers a compelling framework for data scientists to incorporate practices that will '*create more just, equitable, and liveable futures*' (D'Ignazio and Klein, 2020, p. 6). Rooted in intersectional feminist theory (Collins 1991, p. 18; Crenshaw 1991), the authors posit that we must examine intersecting forms of power and oppression (like race and gender) to address the inequalities in our existing data and data practices, and to push back against harmful notions of neutrality that permeate data science. This framework is delineated by seven core principles (Examine Power, Challenge Power, Elevate Emotion and Embodiment, Re-think Binaries and Hierarchies, Embrace Pluralism, Consider Context, & Make Labor Visible).

Much of the reparative work happening in museum collections is rooted in an intersectional approach, critically examining how white, western, and patriarchal epistemes shape the knowledge in our collections. Because of this parallel, *data feminism* offers a generative conceptual framework. By mapping existing LAM, specifically museum, literature to these seven principles, this review serves as a starting point for future empirical research on the extent to which practitioners are willing to implement critical data practices and their specific limitations when trying to do so.

Methods

This review employed purposeful sampling of LAM literature most relevant to the seven principles explored below, following methods for a qualitative systemic review. Rather than traditional inclusion/ exclusion criteria, a qualitative systematic review relies on an assessment of literature to '*mediate messages*' for a narrative synthesis of a particular phenomenon (Grant and Booth 2009, 38). Here, the phenomenon is the established tenants of *data feminism*, aiming to expand their application and relevance to LAM literature as a foundation for future empirical research. This paper does not exhaustively examine all LAM literature yet provides a conceptual framework for the author's empirical dissertation and bridges the gap between these two scholarly areas.

Museum data work is a niche field, with much of the work being conducted by practitioners who do not publish their data practices. As such, literature about feminist work in museums is scattered, often within larger texts and manuscripts. Other literature was found through sources intended for LAM record-keeping more broadly. Examples of journals that were systematically searched included *Collections: A Journal for Museum and Archives Professionals*; *Journal of Documentation*; and *Cataloging and Classification Quarterly*.

One of the intended narratives of this review is to demonstrate that discussions around data practices are ongoing across various types of museums. Notably, natural history museums, which handle the largest volumes of data, have the most extensive body of literature. Because of this, literature from the natural history domain was often excluded to prevent it from overshadowing other areas of museum data work.

Data feminism in museum collections

Examine power

When examining power, D'Ignazio and Klein (2020) give a variety of points to pay attention to understand the '*matrix of domination*' (Collins, 1990), or the intersectional power dynamics that structure our society. They urge us to look at factors such as who is doing the work and for whom, and the silences and privileges that result from that.

D'Ignazio and Klein in their analysis of this principle state that *who* has access to do the work is important (2020). Historically museum work, even invisible museum work, has only been accessible to a certain type of woman: one who is most likely college-educated, and can afford to accept low or no wages, through either other spousal or other family support. Because of this, museum work like other LAM professions continues to hold '*White-IST*', or white and elitist (Mehra and Gray, 2020) institutional values, overall limiting the intersectional work that is acceptable in these spaces. When speaking of power, there are privileged dynamics of race and class at play as well in museum work.

Museum work is a clear example, however, that women having representation in a workforce does not equate to power or equality (Callihan and Feldman 2018). Museums are a feminized profession, with most of the work being done by women across all divisions (Baldwin and Ackerson, 2017; Sweeney et al. 2022;), and follows similar trends of other '*pink collar*' professions of stagnant pay and limited upward mobility (Howe 1978; Nie 2017). Museum professionals also face pay disparities between gender and race, as white women make 71 cents to the dollar of their male colleagues, while Black and Latina women make 60 cents and 55 cents to the dollar of their white male colleagues (Baldwin and Ackerson 2017).

Challenge power

'*Challenging power requires mobilizing data science to put back against existing and unequal power structures and to work towards more just and equitable futures*' (D'Ignazio and Klein 2020, 53). The authors note that principles like equity and co-liberation are critical within the act of challenging power. Many of the issues mentioned above prompt action from practitioners, community members, and scholars to mitigate the negative impacts of power dynamics. Practitioners are taking a proactive accessioning approach to minimize this in the future (Reilly 2018; Halpern and Berns 2019). Notably, many queer museum workers are spearheading efforts to challenge oppressive structures within these institutions (Middleton and Hagen 2022). Similarly, the Guerilla Girls, an artist-activist group, has a long history of calling out institutions that have gender discrepancies in their collections and exhibition practices (Guerilla Girls 2020). These examples, while happening in museum spaces beyond computing, demonstrate that these conversations around feminism and intersectionality are an interest throughout these institutions and a clear space for future collaboration.

In critical data work, refusal and resistance are clear forms of action (Cifor et al. 2019; Garcia et al. 2022; Sutherland 2023) against harmful data regimes. Resistance is also present in museums to challenge oppressive powers as well. This includes cultural heritage workers resisting mass incorporation of potentially harmful artificial intelligence (AI) technology (Dikow 2023; Foka 2023; Murphy 2024) and a refusal to allow the oppressor's standpoint to permeate collection data (Caswell 2022; Patin et al. 2021). From a community standpoint, it also involves a refusal to do things in a way that is convenient to colonial institutions. Robin Gray calls for Indigenous community members to embrace the concept of rematriation, a practice rooted in Indigenous, feminist values that aims for repair beyond repatriation. She states, '*Rematriation involves both a turn away from the colonial order of things and a turn toward Indigenous nationhood*' (Gray 2022). Other examples of this include museums that purposefully centre their data and exhibitions centre around the experiences of their community members, rather than object data (Duhamel and Peristerakis 2017).

Elevate emotion and embodiment

'*Data feminism teaches us to value important forms of knowledge, including the knowledge that comes from people as living feeling bodies in the world*' (D'Ignazio and Klein 2020, 73). Within this principle, *Data Feminism* authors are speaking specifically to the power of bringing emotion and embodiment into our data and pushing back against Donna Haraway's '*god trick*' of data having one singular point of view or seeing '*everywhere from nowhere*' (Haraway, 1988, 581).

Museums have caused real harm to communities, partially through the ways they have decontextualized many of their objects, severing them from any sort of subjective experience. From an exhibition standpoint, many efforts are underway to build affect and emotion into exhibits (Arnold-de Simine, 2013; Katrine, 2017; Kidd, 2018; Smith, 2021), often with the assistance of technology (Benford et al., 2022). The disconnect here is that many museums leave '*interpretation*' up to curatorial and exhibition staff, while those working in museum computing are still dealing with the remnants of the '*old museum*' data (Cameron and Robinson, 2007).

Many recognize this gap (Cameron 2010; Canning 2018; Losh et al 2016; Prosper 2024; Turner 2020); however, this sort of work often works directly in opposition to the narrative format (Manovich 2002) of structured databases, a persistent problem. Some of the most successful examples are when researchers outside of museums take museum data and transform it in embodied and affective ways. (for example: see projects like the *Great Lakes Research Alliance for Aboriginal Arts and Culture* (Bohaker et al., 2014; Rayburn et al., 2024; Wilmott et al., 2016), and projects built on the Mukurtu content management system (Christen, 2017; Christen, 2019; Passamaquoddy Peoples' Digital Archive, n.d; Plateau Peoples' Web Portal, n.d.). However, since much of this work is happening within individual projects, those working in museum computing have not incorporated this into their work practices on a field-wide level.

Re-think binaries and hierarchies

'*Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression*' (D'Ignazio and Klein 2020, 97). The authors also note this work includes examining classification systems for what knowledge or types of data the system intentionally leaves out or does not collect. This principle speaks most directly to the classification schemas used to organize 1) the physical collections in these institutions and 2) their accompanying data that has been produced over time. Museum catalogs and records, for many peoples, have been tools of erasure that reproduce harmful categorizations (Patin et al., 2020), such as referring to Indigenous communities by terms that were imposed upon them by settlers (Phillips, 2011; Turner, 2020). Museum documentation has also been highly skilled at legitimizing oppressive epistemologies like white supremacy and patriarchy into its collection knowledge.

In the last 50 years, there has been a shift in LAM institutions, recognizing the ways that classification systems reinforce systemic oppression. Sanford Berman, one of the first '*radical librarians*' widely recognized to push back against these practices, interrogated Library of Congress Subject Headings in the 1970s (LCSH) (Berman, 1971). His work led to a larger call for action from LAM practitioners beyond libraries, which came to be known as radical cataloging (Roberto 2008), and more recently critical cataloging (Adler, 2017; Snow and Dunbar, 2022; Watson, 2020). Specifically, there has been a focus on creating collection metadata that is inclusive of gender (Billey and Drabinski, 2019; Watson et al., 2023), centring Indigenous epistemes (Cherry and Mukunda, 2015; Duarte and Belarde-Lewis, 2015; Swanson, 2015), and general push back against '*neutral*' information organization systems (Patin et al., 2021). Further, in museums, this work is often referred to as re-description, which in most connotations is directly working against past harmful data schemas (Adler, 2016; Christen and Anderson, 2019; Haberstock, 2020).

Embrace pluralism

This principle of *data feminism* scrutinizes the myth of '*clean data*' and urges data scientists to consider how data cleaning, in many ways, is an act of 'scrubbing' data into a single point of view. D'Ignazio and Klein state, '*data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experimental ways of knowing*' (2020, 125).

Calls for plurality of data in museum records emphasize the need to embrace diverse interpretations of collection items and connect them to various '*cultural, social, historical, technological, artistic, and disciplinary contexts*' (Cameron 2010, 89). Additionally, there are calls to prioritize Indigenous (Duarte and Belarde-Lewis 2015) and Black feminist (Patin et al. 2021) ontologies. Bringing language plurality in records is a common example of this (Castle 2021; Christophe 2024; Lewis 2024; Trammell 2019). Specific examples include the Keaukānaʻi research group partnering with the British Museum to integrate a Hawaiian language taxa into the museum database (Christophe 2024) or the integration of the Sm'algya̲x language at the Burke Museum of Natural History and Culture (Castle 2021). Much of this work also happens with external research and community groups use museum data as a starting point and build plurality into it.

Within this principle, D'Ignazio and Klein also note that when moving away from a single objective truth, uncertainties and ambiguity appear in one's data. Part of embracing plurality is making sure points of uncertainty are visible to users. Uncertainties within museum data are pervasive because of the convoluted and hasty nature that most collections were acquired under. Further, details that were important at the time of collection centered Western and white priorities while intentionally erasing other ways of knowing (Patin et al. 2020). These uncertainties are still present within the data, and in fact, incomplete object records are seen as the norm. Many practitioners are grappling with how to make uncertainties more prominent, rather than let them disappear within the documentation infrastructure. Some practitioners have noted that the task of creating robust documentation should be a museum-wide effort, involving the efforts of curatorial and exhibition staff as well (Mulkerin 2013). In the natural science data, it has been argued that noting places of uncertainty (Faniel et al. 2019; Vollmar et al. 2010) makes collection data stronger and more accurate for future re-use.

Rather than trying to fill in the gaps, other scholars have argued for using points of uncertainty and silence to write '*narrative[s] of what might have been or could have been*' (Hartman 2008, 12). Uncertainty can also be used as an advantage during decolonial work. Alírio Karina, speaking on the uncertainty of object records in the Smithsonian's African ethnographic collections states, '*Uncertain objects…which form a significant portion of African ethnographic collections – offer an opportunity to think outside of a fixing imperial gaze*' (Karina 2020,18). These brief examples demonstrate the multitude of ways memory institutions and those using their collection data are grappling with the idea of uncertainty, but the consensus is that it cannot be ignored.

Consider context

The 6th principle of data feminism urges us to consider the context of the data we use and question how it came to be. In a museum context, this includes who collected the original data, for what purpose, and what data-cleaning efforts have occurred since the original documentation. Hannah Turner, speaking specifically on the colonial legacies of ethnographic collections states, '*when museum documentation is seen as neutral…we must question what the origins of this objectivity are, where it comes from, and how it stands to last*' (2020, p.5). Furthermore, she argues that data cannot be detached from its repositories and their initial aims for collections. Many museums were founded from private collections or '*cabinets of curiosity*' (Strassor, 2012) and later aggregated into public institutions. That means both the collection biases and the data associated with them are in many ways limited by these early collecting efforts. Many anthropological and natural history collections were systemically created with specific calls out to scientists and government officials

for desired objects and species (Parezo, 1987; Turner, 2020); their documentation practices followed a similar positivist methodology. Finally, museums have a long history of helping to legitimize nation-state efforts in colonization and genocide, often through their documentation (Onciul, 2015). This context should also not be overlooked.

Make labor visible

'*The world of data science, like all work, is the work of many hands. Data feminism makes this labor visible so that it can be recognized and valued*' (D'Ignazio and Klein, 2020, p.173). This principle centres around the fact that much data work, especially within large tech companies is unknowingly obtained (Morreale et al., 2023), or the labor of people of colour that are drastically underpaid and purposefully rendered invisible (Gray and Suri, 2019). D'Ignazio and Klein also note that the labor of data entry is dramatically undervalued, a trend most similar to what has historically and presently been the case in museums.

Museums have a long history of undervaluing data entry and rendering this work invisible. When computers were first integrated into museum collections in the 1960s, many museums hired women trained as '*key punchers*' to transcode collection data into the digital medium. This work was viewed as unskilled labor that came with its own industry standard, not as a component of the knowledge production process (Lindsay, 1968). Connotations of data entry work as unskilled have been perpetuated, with consistent trends of hiring students or unpaid interns to do data entry or attaching precocious funding to it through short-term positions (Bryant, 2022; Südkamp, 2021).

Invisibility also relates to the original purpose of museum records. They were created for internal purposes, not meant to be shared with the public. This placement of museum data as part of the museum infrastructure meant that by design, the documentation (and as a result those creating and maintaining the documentation) are designed to fade into the background of the environments they operate within (Bowker and Star, 2000). As such, rarely are museum records '*authored*'. This amplifies the bias problems explored above as it is often unclear who did the documentation under what circumstance, similar to '*black boxing*' that happens in data science. Some museums are trying to work against this by considering the knowledge work that goes into museum computing and are considering ways to author museum records.

A noticeable trend regarding museum labor is the push by practitioners to recognize that museum data work is contingent and low-paid (Bryant, 2022; Rodriguez et al., 2019; Südkamp, 2021). This follows trends in other memory institution work, highlighting the high rates of burnout and exploitation in jobs that have been framed as a '*sacred calling*' (Ettarh, 2018) like librarianship. In museums, especially since the Covid-19 pandemic we have seen mass unionization efforts (Horowitz, 2020; Middleton and Hagen, 2022; Ripley, 2023). The parallel efforts for work visibility are a valuable intersection that data science and museum studies could benefit from further cross-discipline collaboration.

Conclusion

Data feminism is a powerful framework with the potential for significant impact in LAMs, particularly in museums. This review begins to map the principles of data feminism to existing work in LAMs, revealing a longstanding awareness of the limitations inherent in the data schemas that structure collection data. However, it also highlights the need for museums to avoid over-reliance on external researchers and community members to add plurality and embodiment to their data. There is a clear necessity for practical data guidelines for museum practitioners who actively resist harmful notions of objectivity from the past. Additionally, more empirical research is needed to assess the extent to which critical data work is currently being integrated into museum collections, especially as data-driven practices and the use of robust AI systems and large language models (LLMs) continue to grow. Understanding how practitioners can resist harmful tech, and the power structures embedded within these systems will be increasingly vital.

Acknowledgments

This work is partially funded by the University of Michigan museum studies program. Many thanks to my advisors Andrea Thomer and Ricky Punzalan for their continued support of this work.

About the author

Alexandria Rayburn is a Doctoral candidate at the University of Michigan School of Information, where she also received her master's degree in Information Science. She conducts qualitative research on topics related to computing in museums and knowledge infrastructures. Specifically, she studies the data practices and labor required to maintain these systems. She has published in archival science, science technology and human values, and more. She can be contacted at arayburn@umich.edu.

References

Adler, M. (2016). The Case for Taxonomic Reparations. Knowledge Organization 43(8): 630–40. https://doi.org/10.5771/0943-7444-2016-8-630.

Adler, M. (2017). Classification Along the Color Line: Excavating Racism in the Stacks. Journal of Critical Library and Information Studies 1(1). https://doi.org/10.24242/jclis.v1i1.17.

Arnold-de S.S. (2013). Mediating Memory in the Museum: Trauma, Empathy, Nostalgia. Palgrave Macmillan Memory Studies. Houndmills, Basingstoke, Hampshire New York, NY: Palgrave Macmillan.

Baldwin, J.H., and Ackerson, A.W. (2017). Women in the Museum: Lessons from the Workplace. First edition. Abingdon, Oxon; New York, NY: Routledge.

Benford, S., Løvlie, A.S., Ryding, K., Rajkowska, P., Bodiaj, E., Darzentas, D.P., Cameron, H., Spence, J., Egede, J., & Spanjevic, B. (2022). Sensitive Pictures: Emotional Interpretation in the Museum. Conference on Human Factors in Computing Systems, 1–16. New Orleans LA USA: ACM. https://doi.org/10.1145/3491102.3502080.

Berman, S. (1971). Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People. Scarecrow Press.

Bowker, G. C., & Star, S.L. (2000). Sorting Things out: Classification and Its Consequences. Inside Technology. Cambridge, Massachusetts: The MIT Press.

Bryant, H. (2022). Digital Heritage and Contingent Labour in the Museum. Museum International 74(1–2): 112–19. https://doi.org/10.1080/13500775.2022.2157569.

Callihan, E., and Feldman, K. (2018). Presence and Power: Beyond Feminism in Museums. Journal of Museum Education 43(3): 179–92. https://doi.org/10.1080/10598650.2018.1486138.

Cameron, F. (2010). Museums Collections, Documentation and Shifting Knowledge Paradigms. In Museums in a Digital Age, edited by Ross Parry, 80–95. Leicester Readers in Museum Studies. London ; New York: Routledge.

Cameron, F., & Robinson, H. (2007). Digital Knowledgescapes: Cultural, Theoretical, Practical, and Usage Issues Facing Museum Collection Databases in a Digital Epoch. In Theorizing Digital Cultural Heritage: A Critical Discourse, edited by Fiona Cameron and Sarah Kenderdine, 165–91. The MIT Press. https://doi.org/10.7551/mitpress/9780262033534.001.0001.

Canning, E. (2018). Affective Metadata for Object Experiences in the Art Museum.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M. (2020). The CARE Principles for Indigenous Data Governance. Data Science Journal 19 (November): 43. https://doi.org/10.5334/dsj-2020-043.

Castle, B. (2021). Living Context: Indigenous Language Integration at the Burke Museum of Natural History and Culture. http://hdl.handle.net/1773/49505.

Caswell, M. (2019). Dusting for Fingerprints: Introducing Feminist Standpoint Appraisal. Journal of Critical Library and Information Studies 3(2). https://doi.org/10.24242/jclis.v3i2.113.

Cifor, M., Garcia, P., Rault, J., Sutherland, Chan, A.S., Rode, J., Hoffman, A.L., Salehi, N., & Nakamura, L. (2019). Feminist Data Manifest-NO. https://web.archive.org/web/20250101033903/https://www.manifestno.com/

Christen, K., Merrill, A., & Wynne, M.(2017) A Community of Relations: Mukurtu Hubs and Spokes, D-Lib Magazine 23, no. 5/6. https://doi.org/10.1045/may2017-christen.

Christen, K., & Anderson, J.(2019). Toward Slow Archives. Archival Science 19(2): 87–116. https://doi.org/10.1007/s10502-019-09307-x.

Clover, D.E., & Sanford, K. (2016). Knowing Their Place: Feminist Understandings and Perceptions of Women Adult Educators in Museums. Journal of Adult and Continuing Education 22(2): 117–33. https://doi.org/10.1177/1477971416672320.

Collins, P. H. (1990). Black feminist thought in the matrix of domination. Black feminist thought: Knowledge, consciousness, and the politics of empowerment, 138(1990), 221-238.

Christophe, A., Solomon, N. H., Kneubuhl, H., Donnellan, V., & Caldeira, L. (2024). Hosting and Integrating a Hawaiian Language Taxonomy in the British Museum's Collection Database. Collections, 20(1), 96-121. https://doi-org/10.1177/15501906241234945

Crenshaw, Kimberle. 1991. "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color." Stanford Law Review 43 (6). https://doi.org/10.2307/1229039.

D'Ignazio, C., and Klein, L.F. 2020. Data Feminism. Strong Ideas Series. Cambridge, Massachusetts: The MIT Press.

Dikow, R.B., DiPietro, C., Trizna, M., BredenbeckCorp, H., Bursell, M., Ekwealor, J.T.B, Hodel, R. (2023). Developing Responsible AI Practices at the Smithsonian Institution. Research Ideas and Outcomes. https://doi.org/10.3897/rio.9.e113334.

Duarte, M.E., & Belarde-Lewis, M. (2015). Imagining: Creating Spaces for Indigenous Ontologies. Cataloging & Classification Quarterly 53(5–6): 677–702. https://doi.org/10.1080/01639374.2015.1018396.

Ettarh, F. (2018). Vocational Awe and Librarianship: The Lies We Tell Ourselves. In The Library With The Lead Pipe, January. https://web.archive.org/web/20241113050956/https://www.inthelibrarywiththeleadpipe.org/2018/vocational-awe/

Faniel, I. M., Frank, R.D., & Yakel, E. (2019). Context from the Data Reuser's Point of View. Journal of Documentation 75(6): 1274–97. https://doi.org/10.1108/JD-08-2018-0133.

Foka, A., Eklund, L. Løvlie, A.S., and Griffin, G. (2023). Critically Assessing AI/ML for Cultural Heritage: Potentials and Challenges. In Handbook of Critical Studies of Artificial Intelligence,

edited by Simon Lindgren, 815–25. Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00082.

Garcia, P., Sutherland, T., Salehi, N., Cifor, M., & Singh, A. (2022). No! Re-Imagining Data Practices Through the Lens of Critical Refusal. Proceedings of the ACM on Human-Computer Interaction 6 (CSCW2): 1–20. https://doi.org/10.1145/3557997.

Grant, J.M. and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information and Libraries Journal. https://doi.org.10.1111/j.1471-1842.2009.00848.x

Gray, M.L., and Suri, S. (2019). Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Houghton Mifflin Harcourt.

Gray, R.R. (2022). Rematriation: Ts' msyen Law, Rights of Relationality, and Protocols of Return. Native American and Indigenous Studies 9(1): 1-27. https://doi.org.10.1353/nai.2022.0010

Greenwald, D. (2021). Painting by Numbers: Data-Driven Histories of Nineteenth-Century Art. Princeton: Princeton University Press.

Guerrilla Girls. (2020). Guerrilla Girls: The Art of Behaving Badly. 4. Nachdruck. San Francisco, CA: Chronicle Books.

Haberstock, L. (2020). Participatory Description: Decolonizing Descriptive Methodologies in Archives. Archival Science 20(2): 125–38. https://doi.org/10.1007/s10502-019-09328-6.

Halpern, J., and Burns, C. (2019). Museums Claim They're Paying More Attention to Female Artists. That's an Illusion. Artnet. https://news.artnet.com/womens-place-in-the-art-world/womens-place-art-world-museums-1654714.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. Feminist Studies 14 (3): 575. https://doi.org/10.2307/3178066.

Hartman, Saidiya. 2008. "Venus in Two Acts." Small Axe 12 (2): 1–14.Horowitz, Y. A. E. L. (2020). A labor of love, but labor nonetheless. The Journal Theory and Practice, 3.

Howe, L.K. (1978). Pink Collar Workers: Inside the World of Women's Work. New York: Avon.

Karina, A. (2020). Uncertain Objects and Ethnographic Possibilities: Thinking through the Smithsonian-Universal African Expedition. Safundi 21(1): 1–29. https://doi.org/10.1080/17533171.2019.1681175.

Duhamel, K.R., & Peristerakis, J. (2017). Indigenous Women, Intersectionality and Activism at the Canadian Museum for Human Rights, in Feminism and Museums, 345-369.

Kidd, Jenny. 2018. "'Immersive' Heritage Encounters." The Museum Review 3 (1). https://themuseumreviewjournal.wordpress.com/2018/04/16/tmr_vol3no1_kidd/.

Leslie, M.B. (2009). Challenging Science as Usual: Women's Participation in American Natural History Museum Work, 1870–1950. Journal of Women's History 21(2): 11–38. https://doi.org/10.1353/jowh.0.0076.

Lewis, K. (2024). Toward Centering Indigenous Knowledge in Museum Collections Management Systems. Collections: A Journal for Museum and Archives Professionals 20 (1): 27–50. https://doi.org/10.1177/15501906241234046.

Lindsay, K.C. (1968). Computer Input Form for Art Works: Problems and Possibilities. In Computers and Their Potential Applications in Museums: A Conference. New York: Arno Press.

Lowry, James. 2023. 'The Record as Command.' In Archives: Power, Truth, and Fiction, edited by Andrew Prescott and Alison Wiggins. Oxford Twenty-First Century Approaches to Literature. London New York: Oxford University Press.

Madsen-Brooks, L. (2009). Challenging Science as Usual: Women's Participation in American Natural History Museum Work, 1870–1950. Journal of Women's History 21(2): 11–38. https://doi.org/10.1353/jowh.0.0076.

Manovich, L. (2002). The Language of New Media. 1st MIT Press pbk. Ed. Leonardo. Cambridge, Mass: MIT Press.

McTavish, L. (2008). Strategic Donations: Women and Museums in New Brunswick, 1862-1930. Journal of Canadian Studies 42(2): 93–116. https://doi.org/10.3138/jcs.42.2.93.

Mehra, B., and Gray, L. (2020). An 'Owning Up' of White-IST Trends in LIS to Further Real Transformations. The Library Quarterly 90(2): 189–239. https://doi.org/10.1086/707674.

Middleton, M., & Hagen, J.J. (2022). Worker-Led Feminist Mobilizing for the Museum of the Future. European Journal of Women's Studies 29(4): 593–617. https://doi.org/10.1177/13505068221125624.

Morreale, F., Bahmanteymouri, E., Burmester, B., Chen, A., & Thorp, M. (2023). The Unwitting Labourer: Extracting Humanness in AI Training. AI & SOCIETY. https://doi.org/10.1007/s00146-023-01692-3.

Mulkerin, M, S. (2013). Elephants Never Forget, but Museums Do: Investigating Under-Documented Objects in Museum Collections. Collections: A Journal for Museum and Archives Professionals 9(2): 149–66. https://doi.org/10.1177/155019061300900203.

Murphy, O. 2024. Power, Data, and Control AI in the Museum. In AI in Museums: Reflections, Perspectives, and Applications, edited by Sonja Thiel and Johannes Christian Bernhardt.

Nie, T. (2017). 'Far Too Female': Museums as the New Pink-Collar Profession - An Introductory Analysis of Pay Inequity within American Art Museums. Seton Hall University. https://scholarship.shu.edu/dissertations/2315/.

Onciul, B. (2017). Museums, Heritage, and Indigenous Voice: Decolonising Engagement. Routledge Research in Museum Studies 10. London: Routledge.

Parry, R. (2007). Recoding the Museum: Digital Heritage and the Technologies of Change. Museum Meanings. London ; New York: Routledge.

Parezo, N. (1987). The Formation of Ethnographic Collections: The Smithsonian Institution in the American Southwest. Advances in Archaeological Method and Theory 10: 1–47.

Passamaquoddy Peoples' Digital Archive. n.d. About Us. Passamaquoddy Peoples' Digital Archive. Accessed April 7, 2024. https://passamaquoddypeople.com/about.

Patin, B., Oliphant, T., Allard, D., Gray, L., Clarke, R.I., Tacheva J., & Lar-Son, K. (2021). At the Margins of Epistemology: Amplifying Alternative Ways of Knowing in Library and Information Science. Proceedings of the Association for Information Science and Technology 58(1): 630–33. https://doi.org/10.1002/pra2.515.

Patin, B., Sebastian, M., Yeon, J., & Bertolini, D. (2020). Toward Epistemic Justice: An Approach for Conceptualizing Epistemicide in the Information Professions. Proceedings of the Association for Information Science and Technology 57(1): e242. https://doi.org/10.1002/pra2.242.

Phillips R.B. (2011) The digital (R)evolution of museum-based research. Museum pieces: toward the indigenization of Canadian museums. McGill-Queen's University Press, Montreal, 277–296.

Plateau Peoples' Web Portal. n.d. About Us. The Plateau Peoples' Web Portal. Accessed April 7, 2024. https://plateauportal.libraries.wsu.edu/.

Prosper S., Rayburn A.J., Ramirez Y., Punzalan R.L. (2024). Indigenous Digital Projects: A Proposed Assessment Framework. Journal of Information & Culture 59(1): 66-96. https://doi.org/10.7560/IC59104

Rayburn, A.J., Punzalan, R.L. and Thomer, A.K. 2024. "Persisting through Friction: Growing a Community Driven Knowledge Infrastructure." Archival Science 24 (1): 61–82. https://doi.org/10.1007/s10502-023-09427-5.

Reilly, M., & Lippard L.R. (2018). Curatorial Activism: Towards an Ethics of Curating. New York: Thames & Hudson.

Ripley, A.T. (2023). 'Not Just for Coal Miners': Unionization in U.S. Art Museums. Curator: The Museum Journal 66(4): 609–27. https://doi.org/10.1111/cura.12574.

Sepkoski, D. (2013). Towards 'A Natural History of Data': Evolving Practices and Epistemologies of Data in Paleontology, 1800–2000. Journal of the History of Biology 46(3): 401–44. https://doi.org/10.1007/s10739-012-9336-6.

Smith, L. 2021. Emotional Heritage: Visitor Engagement at Museums and Heritage Sites. Museum and Heritage Studies. London New York: Routledge, Taylor & Francis Group.

Snow, K., & Dunbar, A.W. (2022). Advancing the Relationship between Critical Cataloging and Critical Race Theory. Cataloging & Classification Quarterly 60 (6–7): 646–74. https://doi.org/10.1080/01639374.2022.2089936.

Strasser, B.J. (2012). Data-Driven Sciences: From Wonder Cabinets to Electronic Databases. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43(1): 85–87.

Südkamp, C.M. (2021). Personal Sacrifice for the Public Good: Precarious Work at the Museum. https://doi.org/10.17615/f19v-c271.

Sweeney, L., Harkins, D. & Dressel, J. (2022). Art Museum Staff Demographic Survey 2022. Ithaka S+R. https://doi.org/10.18665/sr.317927.

Trammell, K. (2019). The Indigenous Arts Archive: Indigenizing the Spencer Museum of Art's Database. http://hdl.handle.net/1808/27947.

Turner, H. (2020). Cataloguing Culture : Legacies of Colonialism in Museum Documentation. UBC Press.

Walter, M., Lovett, R., Maher, B., Williamson, B., Prehn, J., Bodkin-Andrews, G., &Lee, V. (2021). Indigenous Data Sovereignty in the Era of Big Data and Open Data. Australian Journal of Social Issues 56(2): 143–56. https://doi.org/10.1002/ajs4.141.

Watson, B.M. 2020. 'There Was Sex but No Sexuality*:' Critical Cataloging and the Classification of Asexuality in LCSH. Cataloging & Classification Quarterly 58(6): 547–65. https://doi.org/10.1080/01639374.2020.1796876.

Watson, B. M., Murphy, D., Schaefer, B. & Huang, J. (2023). 'Our Metadata, Ourselves': The Trans Metadata Collective. Proceedings of the Association for Information Science and Technology 60(1): 433–41. https://doi.org/10.1002/pra2.801.

Willmott, C.A., Taitt A., Corbiere M.A., Corbiere AO (2016) Toward language in action: agency-oriented application of the GRASAC Database for Anishinaabe language revitalization. Museum Anthropology Review 10(2):91–116. https://doi.org/10.14434/mar.v10i2.19322

Vollmar, A, Macklin, J.A., & Ford, L. (2010). Natural History Specimen Digitization: Challenges and Concerns. Biodiversity Informatics 7(2). https://doi.org/10.17161/bi.v7i2.3992

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

A semantic framework for digital archives of ancient Chinese villages aimed at humanistic discovery: a case study of four villages in Henan

Tianjiao Qi, Yiduo Wang, Meiman Wong, Tsamchu Basang, and Linglu Li

DOI: https://doi.org/10.47989/ir30iConf47326

Abstract

Introduction. Ancient Chinese villages are crucial cultural legacies facing preservation challenges. Current archiving has drawbacks. This research aims to create a semantic framework for digital archives to boost humanistic knowledge discovery.

Method. We examined three common archiving rules, extracted cultural elements, and designed a semantic framework by reusing international ontologies CIDOC-CRM. Then, we applied it to four villages in Henan for dataization.

Analysis. After comparing the archiving rules and integrating cultural elements, we evaluated the framework. We also processed the villages' archives, including data annotation and Protege knowledge base construction.

Results. The framework consists of 6 entities, 29 relations, and 94 attributes. The semantic structures of the four villages' digital archives were presented, confirming the framework's practicality and applicability.

Conclusion(s). The semantic framework offers a feasible solution for ancient village archive dataization and knowledge discovery. Future efforts should center on enhancing knowledge presentation and utilization for better cultural heritage protection.

Introduction

Ancient Chinese villages, typically referring to settlements established before the Republic of China era, are also known as '*traditional villages*' and '*constitute a significant legacy of Chinese agrarian civilization*' (Feng Jicai, 2012). Renowned for their historical depth and abundance, these villages preserve a wealth of historical information and cultural heritage. As President Xi Jinping has emphasized, '*rural civilization is the cornerstone of Chinese civilization, villages are its carriers, and agrarian and scholarly traditions are our soft power*' (Xin Huashe,2023). However, during China's rapid modernization and urbanization, these ancient villages and their associated traditional cultures have faced critical challenges.

To ensure their preservation, the China Federation of Literary and Art Circles initiated a nationwide survey and archiving of ancient villages in 2003. Since 2012, multiple government departments have jointly launched the Traditional Village Survey and established an expert committee to compile the Traditional Village Directory. By 2023, six batches comprising 8,155 villages of significant conservation value had been included in the directory. Inspired by these national efforts, several provinces, including Anhui and Zhejiang, have completed local '*thousand-village archiving*' projects. These initiatives have yielded substantial progress in documenting ancient villages, laying the groundwork for digital archiving, which has become a focal area in recent years.

> *By the end of 2024, the 'Chinese Traditional Village Museum' had launched digital archives for 1,458 village-specific exhibits*' (Traditional Chinese Village Digital Museum, 2024), '*while Sichuan Province plans to establish a comprehensive village archive database by 2025*' (Luo Zhiyang, 2019).

Similarly, the Taizhou Archives has included digital archiving of traditional villages in its 14th Five-Year Plan.

Motivated by the progress in archiving practices, Chinese academia has conducted extensive studies on the role of ancient village archives in cultural heritage preservation, resource development, cultural discovery, and historical tracing. Scholars generally agree that village archives should encompass tangible cultural heritage, intangible cultural heritage, historical documents, notable historical events, and influential figures. Researchers such as An Hongqingn (An, H.Q.2017, p.29-30), Xu Xinyun (Xu, X.Y.2017, p.43-48), and Ren Yue (Ren, Y.2018, p.79-84) emphasize the importance of documenting historical culture, material forms, and their transformations, while Man Yi (Man, Y. and Wang, P.2017, p.12-15) advocates for the standardization of archiving practices. He Siyuan (He, S.Y.2017, p.49-53) and Zheng Jinxia (Zheng, J.X et al.2023, p.194-198) further explore the institutional mechanisms and roles of archival departments in traditional village conservation. Tang Qi (Tang, Q.2023, p.66-68) argues that archival preservation should go beyond static records to emphasize the 'activation' of traditional villages.

Despite these efforts, significant challenges remain. The vast scale but relatively low perceived value of ancient village archives, coupled with the lack of attention in both practice and theory to these issues, hinder progress. In practice, the diversity of archiving stakeholders and the absence of unified standards lead to severe issues of multi-source heterogeneity, impeding the digitization of archives. Multi-source heterogeneity refers to the complex origins of village archives, which include government departments, cultural institutions, private organizations, and local villagers, resulting in diverse content structures and formats such as text, images, and audiovisual materials. Additionally, differing understandings and expressions of cultural elements among stakeholders further exacerbate semantic discrepancies. These challenges significantly hinder data mining and the effective utilization of village archives. Theoretically, existing studies rarely focus on how the public leverages cultural knowledge in village archives, nor do they propose a systematic

framework or tools to support the structured storage, semantic analysis, and relational mining of unstructured data.

Against this backdrop, this study leverages existing archiving standards and ontology construction techniques commonly used in cultural heritage research. By integrating the internationally recognized CIDOC-CRM semantic framework (FAN, Q. et al.2023, p.469-482), Event Ontology (Raimond, Y., Abdallah, S,2007), and related narrative ontology studies, this research aims to construct a formalized semantic ontology centered on the cultural connotations and core elements of ancient villages. The goal is to explore a feasible pathway for the digitization, knowledge integration, and discovery of village archives. Using archives from four ancient villages in Henan Province as a case study, this study applies the proposed ontology to data digitization, providing a model for the structured archiving and relational mining of ancient village resources. This work seeks to advance the discovery, understanding, dissemination, and preservation of the humanistic value of ancient Chinese villages.

Cultural elements in existing archiving rules for ancient villages

Before constructing a semantic framework for digital archives of ancient villages, this study analyzed common cultural elements in existing archiving rules widely used in China. This analysis provides a foundation for developing and aligning the vocabulary of the proposed ontology. The selected archiving rules include: the Village Information Form for the Digital Museum of Chinese Traditional Villages issued by the Ministry of Housing and Urban-Rural Development of China (Digital Museum of Chinese Traditional Villages,2018), the Text-Based Archiving Form for Traditional Village Surveys and the Image-Based Registration Form for Traditional Village Surveys from the Chinese Traditional Village Network, and the Classification Scheme for Digital Resources of Gaoqian Houku Village developed by the Digital Memory team at Renmin University of China(National Archives Administration of China,2021). These rules represent key approaches at the levels of national registry, grassroots survey, and digital database construction.

Characteristics and differences in existing archiving rules

The Village Information Form for the digital museum of Chinese traditional villages, published in 2018, is a foundational document for digital archiving of ancient villages. Created by the Ministry of Housing and Urban-Rural Development with support from the China Academy of Urban Planning and Design and local administrative departments, it provides a unified template for documenting traditional villages nationwide. This form emphasizes the systematic recording of basic village information and architectural culture, covering geographic location, historical development, architectural characteristics, and cultural heritage. It highlights detailed documentation of village history, migration processes, significant historical figures, and major events. For cultural heritage, it systematically captures data on architectural elements (names, materials, styles, and functional layouts) and includes non-material aspects such as folklore, festivals, rituals, and local dialects, along with information on cultural activities and inheritors of intangible heritage. Notably, the form also includes data on local cuisine, products, and tourism resources, supporting economic development and cultural tourism initiatives.

The Text-Based Archiving Form and Image-Based Registration Form from the Chinese Traditional Village Network were developed as part of the '*Preserving Nostalgia—Traditional Village Survey and Archiving*' project, launched in 2014 with support from the Ministry of Housing and Urban-Rural Development and various cultural organizations. These guidelines focus on comprehensive documentation of cultural and social life, including social structure, population distribution, customs, production methods, natural and cultural heritage, and historical changes, supplemented by extensive visual materials. The rules emphasize combining textual and visual data to document the natural environment, geological conditions, village morphology, architectural relics, and street information, aiming to capture the interconnectedness of human and natural elements. For

cultural heritage, the guidelines prioritize recording items listed in heritage registries, such as historical artifacts, family histories, and manuscript collections, often supported by photographs to vividly depict traditional customs, production activities, and even village inhabitants.

The classification scheme for digital resources of Gaoqian Houku Village, developed by Renmin University of China, was created under a major National Social Science Fund project focusing on the digital preservation and transmission of historical and cultural villages. Starting in 2016 with Gaoqian Village in Xianju County, Zhejiang Province, the project has emphasized the use of digital technologies for resource classification and association. This framework is technically oriented, facilitating the digitization and dissemination of cultural resources. It incorporates oral history recordings, such as interviews with elder villagers and cultural inheritors, to preserve the living memories of ancient villages. Additionally, it provides guidelines for 3D modeling and digital collection of spatial layouts, natural environments, and architectural structures. The framework is characterized by its focus on historical records, family genealogies, manuscripts, and multimedia documentation, showcasing a professional archival approach.

Cultural elements in existing archiving rules

A comparative analysis of the three archiving rules in terms of data structures, content coverage, and technological applications revealed distinctive characteristics. The Ministry of Housing and Urban-Rural Development's rule emphasizes standardization and comprehensiveness, focusing on detailed basic data for unified national management and systematic conservation. The Chinese traditional village network's guidelines prioritize cultural and social life, combining textual and visual documentation to highlight the depth and breadth of cultural transmission and social structures. Renmin University's framework centers on the structured management and association of digital resources, leveraging modern technologies to enhance the efficiency and dissemination of archival materials.

Despite these strengths, existing rules primarily offer checklist-style guidelines for data collection without fine-grained annotation standards for the cultural, heritage, and memory elements within the archives. This limitation makes it difficult to identify distinct cultural characteristics from archived data, leading to the phenomenon of '*standardized archiving with little distinction*'.

To address this issue, this study identified a core set of cultural elements shared across the three rules, forming the foundational vocabulary for the proposed ontology framework. As illustrated in Figure 1, the study integrates key concepts related to village history, spatial relationships, and human elements. Cultural elements concerning historical narratives include the development history of villages, family lineages, heritage evolution, and architectural history. Spatial elements emphasize natural landscapes, cultural landmarks, spatial layouts, and architectural imagery. Human elements focus on identifying key figures who have significantly influenced the historical and spatial development of ancient villages.

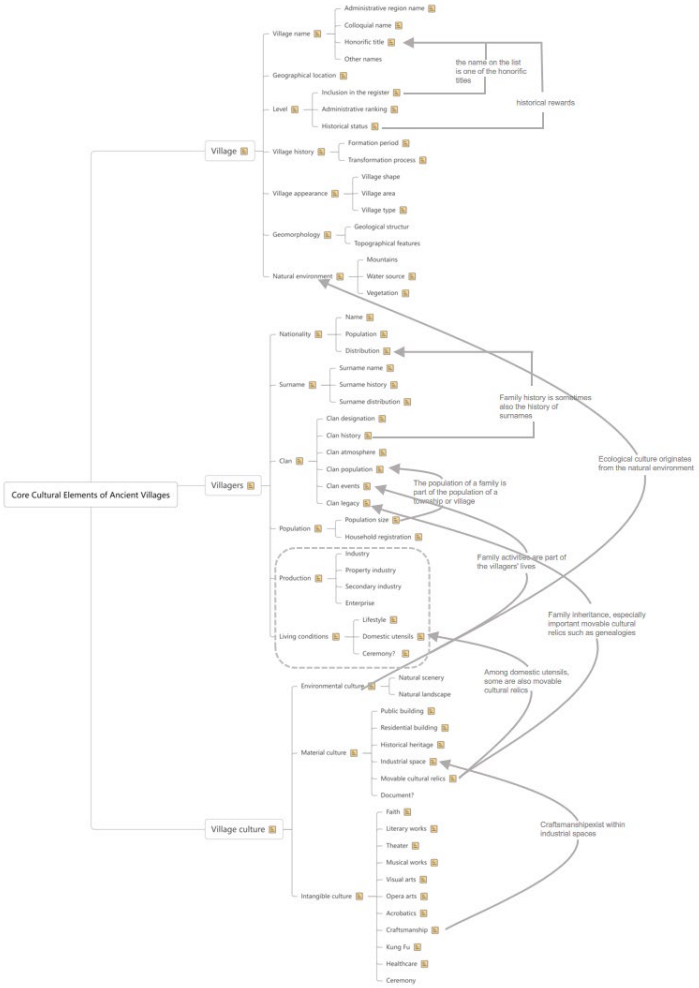


Figure 1. Core cultural elements of ancient villages in existing archiving rules

The systematic identification of core cultural elements of ancient villages addresses the limitations in data annotation within the three major archiving rules. However, it lacks the establishment of diversified relationships between elements, failing to support deeper exploration and discovery of cultural knowledge associated with ancient villages. Moreover, the identified core elements exhibit semantic redundancy or ambiguity when aligned with common semantic frameworks in cultural heritage or historical research. This necessitates further work on semantic disambiguation and alignment to build a systematic and clear semantic model. Such a model would facilitate uncovering humanistic connections within ancient village archives and establish linkages between these archives and other related historical and cultural heritage resources.

Design of a semantic framework for digital archives of ancient villages based on general ontology reuse

After extracting core cultural elements from existing archiving rules, this study further developed a semantic framework for digital archives of ancient villages by reusing general ontologies commonly employed in international cultural heritage and historical research. The CIDOC conceptual reference model (CIDOC-CRM) (Raimond, Y., Abdallah, S.2007) was the primary ontology reused, while the event ontology was applied to describe historical events related to ancient villages. Research on narrative ontologies (Ciotti, F.2016, p.29-44) was utilized to describe the stories of historical figures associated with ancient villages.

Adaptability of general ontologies

The CIDOC-CRM is an internationally recognized conceptual model developed by the International Committee for Documentation of the International Council of Museums (ICOM). It establishes a universal and extensible semantic framework for integrating cultural heritage information, enabling semantic definitions and structured descriptions of concepts and relationships in cultural heritage resources. The model's comprehensive descriptions of entities (e.g., entity) and properties (e.g., property) allow accurate depiction and relational mapping of key aspects such as people, places, and events associated with ancient villages. This makes it well-suited for resource integration and organization.

For this study, the CIDOC-CRM model was adapted to represent the geographical location of villages using the E53 Place category, individuals of historical significance as E21 Person, and relationships such as P8 took place on (linking events to relevant objects) and P74 had current or former residence (indicating places of residence over time). However, the generic nature of CIDOC-CRM presents limitations in addressing the specific characteristics of ancient villages. For example, the E41 Appellation category does not adequately represent the unique cultural designations of villages, which include architectural styles, traditional crafts, natural landscapes, and historical artifacts. Therefore, this study extended the '*appellation*' attribute to describe distinct cultural elements based on traditional archiving rules.

The event ontology provides a structured framework for describing and categorizing events, offering concepts such as event: event, event: factor (elements within an event), and event: product (event outcomes). However, it lacks specificity for relationships commonly seen in the historical events of Chinese ancient villages, such as indirect causal links between village founding or migration and family-related events, or implicit correlations between customs or production habits and wartime events. This study first categorized common historical event types of ancient villages and plans to further analyze complex event relationships in future research.

Existing narrative ontology research, while lacking uniform semantic frameworks, typically describes characters, their attributes and actions, story locations, and related events. These narrative elements complement the CIDOC-CRM and event ontology by filling gaps in the description of narrative aspects, particularly character-related attributes.

Based on the reuse of general ontologies, this study constructed a preliminary semantic framework for ancient villages. The framework was integrated with digital archive data from Gaoqian village in Zhejiang Province, which the research team has been developing since 2016. The framework was evaluated against a classification directory and conceptual model of cultural resources agreed upon by an expert team specializing in ancient villages. Expert feedback indicated that the current model provides high coverage of cultural elements, accurately infers relationships between cultural heritage and historical events, and offers good extensibility for unique features in the natural and cultural environments of ancient villages.

Basic components of the semantic framework for digital archives of ancient villages

Building upon the classification and alignment of core cultural elements, this study expanded descriptions to capture the unique cultural phenomena and attributes of ancient villages. The resulting digital archiving system centers on ancient villages and encompasses seven major entities, 29 relationships, and 94 properties. As shown in the partial entity-relationship (ER) diagram in Figure 2, the semantic framework includes six main categories: basic village information, natural heritage, tangible cultural heritage, intangible cultural heritage, historical documents, historical events, and notable figures.

This framework aims to support in-depth exploration of the historical context, spatial relationships, and human networks of ancient villages, providing a foundation for uncovering and analyzing their rich cultural heritage.

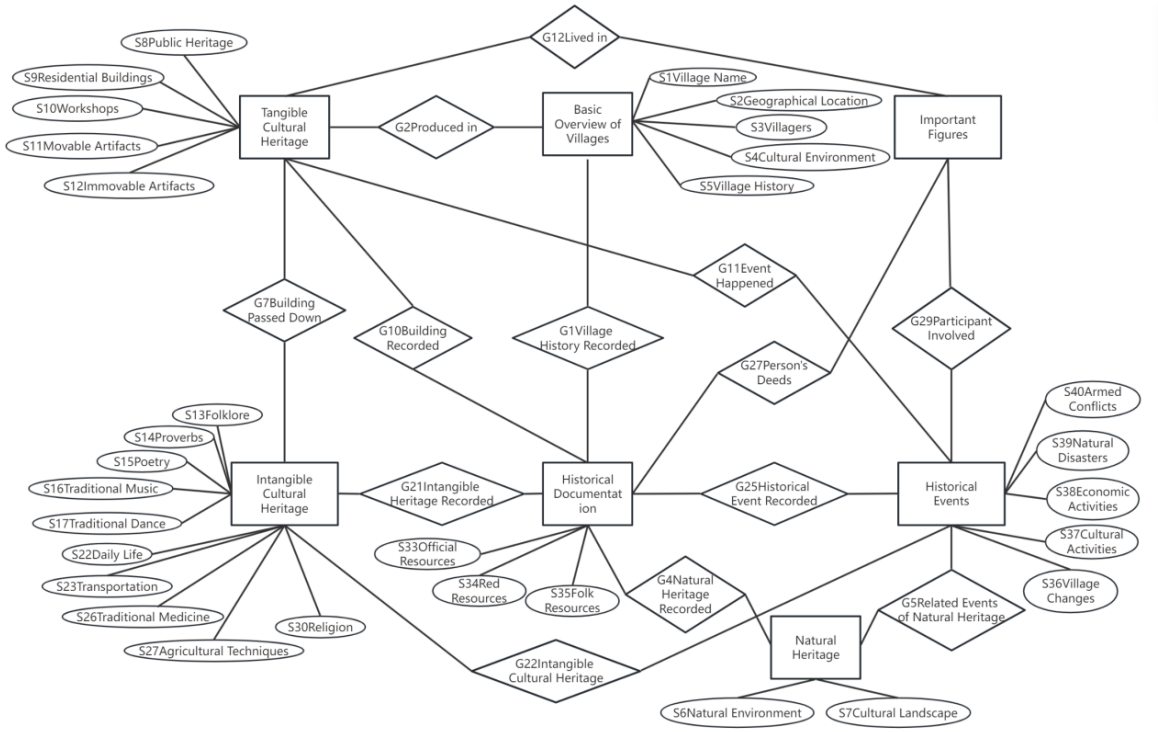


Figure 2. Sketch of digital archiving ER for ancient villages (only some relationships are presented)

Entity types in the semantic famework for digital archives of ancient villages

By analyzing existing rules for archiving ancient villages, it was found that aspects such as village overview, natural environment, and historical and cultural heritage are universally present in all the rules. Specifically, the '*registration form for Chinese traditional villages*' provides the most concise summary, the '*Information Table for Museumization of Chinese Traditional Villages*' offers the most comprehensive classification and content for tangible and intangible cultural heritage, The archiving standards developed by the Digital Memory of Ancient Villages Team at Renmin University of China provide the most detailed descriptions of village overviews and residents. In summary, this study excludes descriptions of tourism guidance in the '*information table for museumization of Chinese traditional villages*', integrating culinary products and agricultural specialties into sections on folk customs, daily life, and knowledge skills. It adopts the maximum union of these three sets of rules and reuses the CIDOC-CRM standards suitable for describing ancient village resources to create a framework encompassing six main entities. (See Table 1).

Entity	Description	Example	Resource Type	Entity Code	Reuse Standard	Remarks

	Village Name		Refers to the administrative name of the village, typically given by the government based on geography, history, and culture	Haoshangqiao Village, Dayu Town, Gongyi City, Zhengzhou, Henan Province		S1		
Basic Overview of Villages	**Geographical Location**		The administrative area where the village is located, including province, city, township, etc.	Dayu Town, Gongyi City, Zhengzhou, Henan Province		S2	E53 Place	
	Villagers		Residents living in the village, including social, cultural, economic, and			S3		
	Cultural Environment		The social and cultural conditions in the village			S4		
	Village History		Historical descriptions of the village, including the formation era, reasons for its		Text records such as village chronicles	S5		
Natural Heritage	**Natural Environment**		The natural geographical conditions of the village			S6		
	Cultural Landscape		Landscapes that combine nature and human activities, holding cultural			S7		
		Public Heritage	Cultural heritage jointly owned, used, and maintained by the community, holding significant historical, cultural, social, or economic value	Temples, ancestral halls, stages, pagodas, historical sites, ancient bridges, wells, tombs, walls		S8		Organized by time, in the format "Building-Time"

Tangible Cultural Heritage	**Architecture**	**Residential Buildings**	Buildings used for daily living purposes	Cave dwellings, courtyards, roundhouses, bamboo houses, stilt houses	Text records such as village chronicles, photos, architectural drawings, and other artifacts	S9		
		Workshops	Places where people engage in handcrafted production and processing	Oil mills, papermaking workshops, tofu factories, dye houses, ceramic workshops, blacksmiths		S10		
	Artifacts	**Movable Artifacts**	Items of cultural heritage that can be transported or moved	Artworks, manuscripts, books, representative objects	Text records such as village chronicles, photos, drawings, and other artifacts	S11		First indicate storage location, then specify
		Immovable Artifacts	Cultural heritage that cannot or should not be moved due to its scale, nature, or association with a	Steles, murals, ruins, temples, city walls		S12		
Intangible Cultural Heritage	**Oral Culture**	**Folklore**	Cultural elements transmitted and preserved orally rather than in written form	Heroic legends, mythological stories		S13		
		Proverbs	Agricultural proverbs, life wisdom, observations of nature, social			S14		
		Poetry	Pastoral poems, labor songs, festival poetry, lyrical poems			S15		
		Traditional	Forms of art expressed through physical actions,	Folk songs, guqin (zither), erhu (two-		S16		

		Music	voice, expressions, and language	stringed fiddle)				
	Performing Arts	**Traditional**	Festive dances, ethnic minority			S17		
		Traditional Drama	Peking opera, Yue opera, Huangmei opera, shadow puppetry			S18		
		Quyi (Chinese Folk Art Forms)	Xiangsheng (comic dialogue), storytelling, drum songs		Text records such as village chronicles, photos, audio files, video files	S19		
		Acrobatics	Tightrope walking, face-changing, contortion			S20		
		Traditional Martial Arts, Sports, and Compet	Taijiquan, Baduanjin (Chinese exercise), dragon boat races, wrestling, archery			S21		
Intangible Cultural Heritage	**Folk Life**	**Daily Life**	Traditional lifestyle of the village, an important part of cultural identity	Farming, folk games, folk arts		S22		
		Transportatio	Horse carriages, water transportation			S23		
		Rituals and Custom	Wedding and funeral customs, sacrificial rituals, banquet			S24		
		Household Beliefs	Ancestor worship, deity worship, feng shui beliefs			S25		
	Knowledge and Skills	**Traditional Medicine**	Knowledge and skills gained through time, experience, and learning in the village	Herbal medicine, acupuncture, cupping therapy, massage therapy		S26		

		Agricultural Techniques	Sowing, fertilizing, irrigation			S27		
		Traditional Handicrafts	Weaving, embroidery, pottery making, carpentry			S28		
		Knowledge of Natural and Astronomical Phenomena	Seasonal changes, star observation, feng shui			S29		
	Spiritual Beliefs	**Religion**	Belief systems collectively followed in the village that influence values, behavior, and	Taoism, Buddhism, Confucianism		S30		
		National Spirit	Solidarity, hard work, frugality, respect for nature, including special regional or wartime			S31		
		Moralit	Filial piety, integrity			S32		
Historical Documentation	**Official Resources**		Documentation, description, or analysis of figures or events related to the village	Village chronicles, local records, land deeds, ration tickets	Text records, photo archives	S33		
	Red Resources		Revolutionary documents, stories of significant figures during revolutionary times			S34		
	Folk Resources		Maps, objects collected from non-official sources, including ancestral portraits,			S35		
Historica	**Village**		Significant events in the village that have	Foundation of the village,	Text records,	S36		

l Events	**Changes**		shaped its development	relocation process	photo archives			
	Cultural Activities		Festivals, religious events, educational activities, cultural exchanges			S37		
	Economic Activities		Agricultural production, handicrafts, commercial trade			S38		
	Natural Disasters		Floods, droughts, earthquakes			S39		
	Armed Conflicts		Events related to defense and resistance during			S40		
Important Figures			Individuals who have made significant contributions or have had a major influence on the	Cultural inheritors, historians		S41	E21 Person	

Table 1. Entity table of the semantic framework for digital archives of ancient villages

Entity relationships in the semantic framework for digital archives of ancient villages

The relationships between entities in the semantic framework for digital archives of ancient villages primarily occur between various types of cultural heritage and historical narratives. Tangible cultural heritage, intangible cultural heritage, and documentary heritage each record, reflect, and reveal the historical events, figures, activities, and phenomena of ancient villages. As shown in Table 2, these relationships not only construct a multidimensional network of cultural resources but also, through interconnected nodes, demonstrate the dynamic evolution of ancient village histories. For example, tangible cultural heritage, such as ancient buildings and sites, is often closely linked to specific historical events or figures. Through the documentation of documentary heritage, the social context and cultural significance behind these events or figures can be further revealed. Intangible cultural heritage, such as traditional crafts and folk activities, when associated with documentary and tangible cultural heritage, reflects the lifestyles and spiritual aspects of the residents of ancient villages. These multidimensional entity relationships not only help to comprehensively understand the cultural significance of ancient villages but also provide a scientific basis for their preservation and inheritance.

Code	Relationship Name	Description	Entity 1	Entity 2	Reuse Standard
G1	Village History Records (Recorded)	Historical documents record the history of the village	Village History	Historical Documents	
G2	Produce in	The place where production takes place	Production	Workshop	
G3	Used for Production (Used for)	Tools used in production	Production	Movable Artifact	

G4	Natural Heritage Records (Recorded)	Historical documents record the natural heritage	Natural Heritage	Historical Documents	
G5	Related Events of Natural Heritage	Historical events related to the natural heritage	Natural Heritage	Historical Event	
G6	Buildings based on skills	The craftsmanship used in building construction	Building	Traditional Handicrafts	
G7	Building Passed Down	Oral traditions related to the building	Building	Oral Culture	
G8	Building Used for (Required)	The application of buildings in folk life	Building	Folk Life	
G9	Building Carries (Recorded in)	The spiritual beliefs carried by the building	Building	Spiritual Belief	
G10	Building Recorded (Recorded)	Historical documents record the building	Building	Historical Documents	
G11	Event Happened (Happened in)	Events that happened in the building	Building	Historical Event	P8 - took place on or within (witnessed)
G12	Lived in (Residence)	The residence where the person lived	Building	Important Person	
G13	Artifacts based on skills	The craftsmanship used in the creation of artifacts	Artifact	Traditional Handicrafts	
G14	Artifact Passed Down	Oral traditions related to the artifact	Artifact	Oral Culture	
G15	Props Used for (Used in)	Props used in performing arts	Artifact	Performing Arts	
G16	Artifact Used for (Required)	The application of artifacts in folk life	Folk Life	Artifact	
G17	Artifact Carries (Recorded in)	The spiritual beliefs carried by the artifact	Artifact	Spiritual Belief	
G18	Artifact Records (Recorded)	Historical documents record the artifact	Artifact	Historical Documents	
G19	Artifact Related Event	Historical events related to the artifact	Artifact	Historical Event	
G20	Records Person (Recorded in)	The artifact records the person's life and contributions	Artifact	Important Person	
G21	Intangible Heritage Recorded (Recorded)	Historical documents record the intangible cultural heritage	Intangible Cultural Heritage	Historical Documents	

G22	Intangible Heritage Related Event	Historical events related to intangible cultural heritage	Intangible Cultural Heritage	Historical Event	
G23	Intangible Heritage Passed Down	Oral traditions related to intangible cultural heritage	Intangible Cultural Heritage	Oral Culture	
G24	Tools Used (Used in)	Tools, decorations used in intangible cultural heritage	Intangible Cultural Heritage	Movable Artifact	
G25	Historical Event Records (Recorded)	Historical documents record the historical event	Historical Event	Historical Documents	
G26	Historical Event Passed Down	Oral traditions related to the historical event	Historical Event	Oral Culture	
G27	Person's Deeds	The main deeds and contributions of the person	Important Person	Historical Documents	
G28	Person Passed Down	Oral traditions related to the important person	Important Person	Oral Culture	
G29	Participants have (Participates in)	Participants involved in the historical event	Important Person	Historical Event	

Table 2. Entity relationships in the semantic framework for digital archives of ancient villages

Attribute descriptions in the semantic framework for digital archives of ancient villages

During the definition of the basic intensions and extensions of the six major entities and their subcategories, this study conducted a general description of the attributes commonly found in ancient villages across various regions and types in China. A basic attribute table was created, listing the attribute descriptions of ancient village entities in detail, including but not limited to geographical location, historical development, architectural style, cultural value. By accurately describing these attributes, it becomes possible to reflect the multidimensional information of ancient villages and provide abundant data support for subsequent mining of digital archive resources more comprehensively. In addition, the definition and classification of attributes refer to the standards of CIDOC-CRM, ensuring consistency and interoperability of data. Through this approach, the semantic analysis of ancient village digital archives can go beyond the level of entities and relationships to delve into the details of attributes, thus providing a solid foundation for future research and applications.

Code	Attribute Name	Entity	Attribute Description	Example	Reuse Standard
D1	Common Name	Village Name	The informal name of the village commonly used by local residents, possibly derived from the village's history, geography, or famous stories	For example, Haishangqiao Village is commonly known as "Haiyan"	
D2	Honorary Title		Honorary titles awarded to the village by the government or other authorities in recognition of its unique cultural, historical, economic, or social contributions	For example, listed in the fifth batch of Chinese traditional villages	

D3	Ethnicity	Villager	The ethnic distribution in the village and related historical activities	For example, Han ethnicity	
D4	Surname		The distribution of surnames in the village and related historical activities	For example, Wang family in Haishangqiao Village	
D5	Clan		The distribution of clans in the village and related historical activities	For example, the Hua family clan in Youdian Village	
D6	Population		The population count and household registration in the village		
D7	Production		The main industries in the village and the villagers' way of life	For example, agriculture, livestock, fishery, forestry, handicrafts, tourism	
D8	Streets and Alleys	Cultural Environment	The distribution of functional areas in the village	For example, residential, commercial facilities, entertainment facilities, village hospital	
D9	Defense Structure		Structures built to protect the village from natural disasters or military attacks	For example, city walls, fortresses, air-raid shelters, dikes	
D10	Roads		The distribution of roads in the village		
D11	Formed Time Is	Village History	Important event nodes in the historical development of the village	For example, Haishangqiao Village was established in the fifth year of the Qianlong reign (1740)	P4 - has time-span (is time-span of)
D12	Formation Reason		The reason for the formation of the village	For example, a natural spring in the western valley led to the construction of a bridge for residents to cross	
D13	Evolution Process		The process of village evolution, including formation, migration, etc.		
D14	Area	Natural Environment	The area of the village	For example, 2865 acres of arable land, approximately 6.7 square kilometers	

D15	Shape		The type and terrain of the village	For example, mountain, hill, plateau	
D16	Vegetation and Animals		The biological coverage of the village	For example, crops like wheat, corn, millet, sweet potatoes, sesame, peanuts, sunflowers, rapeseed	
D17	Soil Type		The type of soil in the village, such as red soil, black soil, yellow soil	For example, red soil, black soil, yellow soil	
D18	Hydrology		The hydrological conditions of the village, such as rivers and lakes	For example, river basins, wetlands	
D19	Mountain and Terrain		The mountains and terrain of the village	For example, backed by the Taihang Mountains, flowing through the Jinsha River	
D20	Climate Type		The climate zone of the village	For example, subtropical climate, temperate monsoon climate	
D21	Natural Disaster		The types of natural disasters experienced by the village	For example, mudslides, floods	P4 - has time-span (is time-span of)
D22	Cultural Landscape Type Is	Cultural Landscape	The type of cultural landscape in the village	For example, ancient trees, ancient wells	
D23	Formation Time Is		The formation time of the cultural landscape		
D24	Maintenance Time Is		The time when the cultural landscape was maintained		
D25	Disappearance Time Is		The time when the cultural landscape disappeared		
D26	Formation Place Is		The place where the cultural landscape was formed		
D27	Cultural Landscape Recorded Has		The records related to the cultural landscape		
D28	Function Has		The main function of the cultural landscape	For example, ancient wells providing freshwater	P101 - had as general use (was use of)

D29	Formation Time Is	Building	The formation time of the building		P4 - has time-span (is time-span of)
D30	Maintenance Time Is		The time when the building was maintained		
D31	Disappearance Time Is		The time when the building was destroyed		
D32	Formation Place Is		The place where the building was formed	For example, the Wang family courtyard in Haishangqiao Village in Zhengzhou, Henan	
D33	Events Happened		Events that took place in the building	For example, France and Germany signed armistice agreements in the same railway carriage in 1918 and 1940	P8 - took place on or within (witnessed)
D34	Area Is		The area occupied by the building		
D35	Number of Floors Is		The number of floors in the building		
D36	Creator Is		The designer or creator of the building		
D37	Maintainer Is		The person who maintains the building		
D38	Resident Is		The past and present residents of the building		
D39	Owner Is		The past and present owners of the building	P52 - Current owner (owns it now)	
D40	Building Type Is		The type of building	For example, ancestral hall, temple, academy, drum tower, opera stage, guild hall	
D41	Protection Level Is		Whether the building is included in various protection lists	For example, national protection, provincial protection, municipal protection, county protection	

Table 3. Attribute table in the semantic framework for digital archives of ancient villages (Partial Illustration)

The humanistic discovery capabilities of the semantic framework for digital archives of ancient villages: a case study of four villages in Henan province

Overview of the four villages in Henan province and their digital archives

Henan Province boasts abundant political, military, artistic, and cultural resources, nurturing ancient cities and preserving numerous historically rich ancient villages. This study selects four villages in Henan: Haishangqiao, Yuanqiao, Youdian, and Dawangmiao —for integrated analysis of their digital archive resources. By conducting semantic analysis based on the semantic framework proposed in this study, the research aims to verify the applicability of the semantic framework to the analysis of actual digital archives of ancient villages and explore innovative paths for constructing and presenting humanistic knowledge associations within ancient villages.

These four villages, all listed in the catalogue of Chinese traditional villages, possess rich archival resources covering almost every aspect of the archiving system, with unique characteristics for each village: Haishangqiao preserves intact archives of notable residences and biographies, featuring a wealth of personal records; Yuanqiao retains plaques from the Ming and Qing dynasties and historic buildings, offering substantial tangible archives; Youdian, known for its distinctive cave dwellings and sunken courtyards, holds abundant property-related archives and Dawangmiao served as the site of Henan University during the anti-Japanese war period and contains significant manuscript archives related to red cultural resources. From the perspective of archival resource types, the digital archives of these four villages are primarily composed of text and image resources, with limited digitalized audio and video resources. In terms of archival content, most of the information concentrates on the past 50 years, with some records of legends and background concerning the villages' founding but lacking detailed textual and physical documentation. The specific conditions of the digital archives of these four villages are presented in Table 4 (see Table 4).

Resource Status		Dayugou Town, Haishangqiao Village, Gongyi City, Zhengzhou, Henan Province	Dajindian Town, Yuanqiao Village, Dengfeng City, Zhengzhou, Henan Province	Shanhua Town, Youdian Village, Yanshi District, Luoyang, Henan Province	Tantou Town, Dawangmiao Village, Luanchuan County, Luoyang, Henan Province
Basic types of digital resources	**Text Resources**	'Haishangqiao Village Chronicles', Wang Family Genealogy, 'Da Jie Stele Inscription' etc.	'Speaking of Yuanqiao', List of attendees at the first Dengfeng County Party Congress, 'Dengfeng County Chronicles', stele inscriptions, etc.	'Hua Family Genealogy', 'Youdian Town's Restoration of the Bull King Temple', 'Hong Xing Guan's Patching Stele', 'Establishment of Jade Pavilion', 'Yan County Chronicles', 'Youdian Village Chronicles', etc.	'Sun Family Genealogy', 'Imperial Edict Stele', 'Dawangmiao Village', Sun Family Genealogy, etc.

	Image Resources	Real estate certificates from the Republic of China period, deeds, food coupons from the early PRC period, historical buildings like Guang ye Wang's residence, etc.	Deeds, property rights, tax payment certificates from the Republic of China period, Yuan Family Genealogy, historical buildings like tunnels, earthen ramparts, San Xian Temple, etc.	Donation lists, historical buildings like cave dwellings, ancient wells, ancient north, ancient plaques, etc.	Route maps and classroom design drawings of Henan University Agricultural College, student IDs, photos of famous scholar Jing run Chen speaking at Henan University, historical buildings like Wen Qu Bridge, Sun Family Mansion, Li Family Mansion, etc.
	Video Resources		Yuanqiao 1979		
	Data Resources			Statistics of scholars from the Ming and Qing dynasties, statistics of college students before the founding of New China, statistics of junior college students after the liberation of Youdian Village, statistics of political, military, educational, and business situations after the liberation of Youdian Village, etc.	
	Online Resources	Traditional villages on the 'Track together' website	Traditional villages on the 'Track together' website	Traditional villages on the 'Track together' website	Traditional villages on the 'Track together' website
Features	**Spatiotemporal coherence**	From the Qianlong era of the Qing Dynasty to the present, the village history	From the Hongwu era of the Ming Dynasty to the present, the village history	From the Northern Song Dynasty to the present, the village history has been documented	From the late Ming and early Qing dynasties to the present, the village history has

	has been recorded in the village chronicles.	has been documented in the book 'Speaking of Yuanqiao'.	in the 'Youdian Village Chronicles'.	been documented in the 'Dawangmiao Village'.
cultural heritage preservation	Well-preserved historical buildings and village civilization	Well-preserved historical buildings and village civilization		The village's red gene from the period of the war of resistance against Japan continues to this day
Storability and replicability	primarily text and image resources	primarily text and image resources	primarily text and image resources	primarily text and image resources
modality	Static	Static	Static	Static

Table 4. Overview of digital archive resources for the four villages in Henan province

Data processing workflow for digital archive resources of the four villages in Henan province

Based on resources such as village chronicles, museum collections, oral accounts from villagers, and field visit photographs, this study first standardized and annotated the data using Excel, following the semantic framework for digital archives of ancient villages described in Section 3.2. Subsequently:

Hierarchical entities were created using Protégé, with relationships and attributes added to incorporate the resources of the four villages. A knowledge base was built in Protege. To ensure clarity, flexibility, and distinctiveness, the OntoGraf visualization plugin was used to present the framework. For instance, Haishangqiao village highlights residential buildings associated with the Wang family, while Yuanqiao emphasizes the relational network of the Yuan family.

Semantic structure presentation of digital archives for the four villages in Henan province

Haishangqiao village, located in Dayugou town, Gongyi city, Zhengzhou, Henan province, is renowned for its historically significant Wang family genealogy and the '*Haishangqiao village chronicle*'. The history of the village's resources can be traced back to the Qianlong era of the Qing dynasty, with textual and image data as the primary materials. The village's historic architectural complex is a remarkable sight, with its overall layout well-preserved. The Wang family residence, as an iconic building, highlights the village's unique cultural characteristics. Within the resource framework, '*residential buildings*' emerge as a central element, radiating outward to form a distinct spatial layout. Detailed records of folk customs and daily life are also fully reflected in the framework, offering subsequent researchers an abundant and meticulous collection of materials.

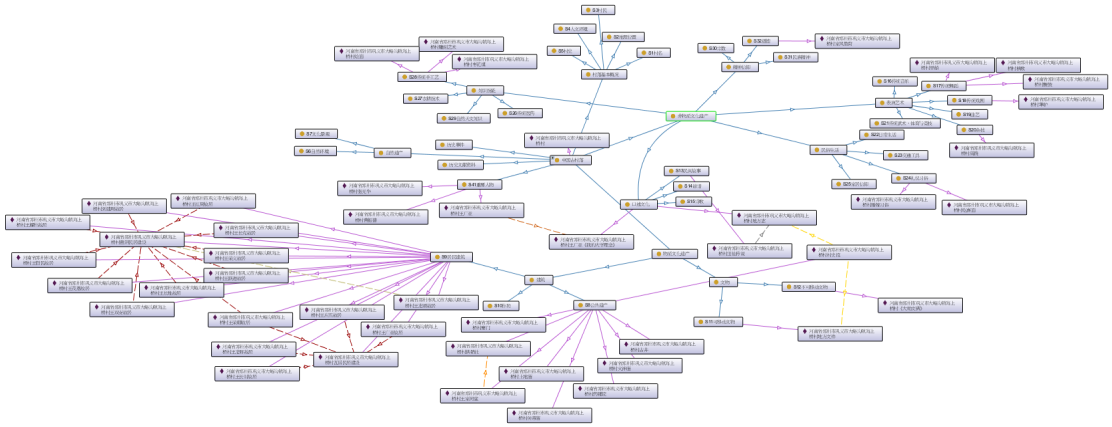


Figure 3. Semantic structures presented in the digital archives of Haishangqiao village

Yuanqiao village, located in Dajindian town, Dengfeng city, Henan province, possesses abundant resources. Centered on the village chronicle '*talking about Yuanqiao*', the archives provide detailed records of biographies, historical relics, events, and folk customs, with its history tracing back to the Hongwu era of the Ming dynasty. Numerous immovable cultural relics, including a significant number of inscribed stone tablets, have been well preserved. In the resource framework diagram, the Yuan family lineage serves as the main thread, clearly illustrating the continuity and evolution of Yuanqiao village in terms of folk culture, village development, historical progression, and economic growth.

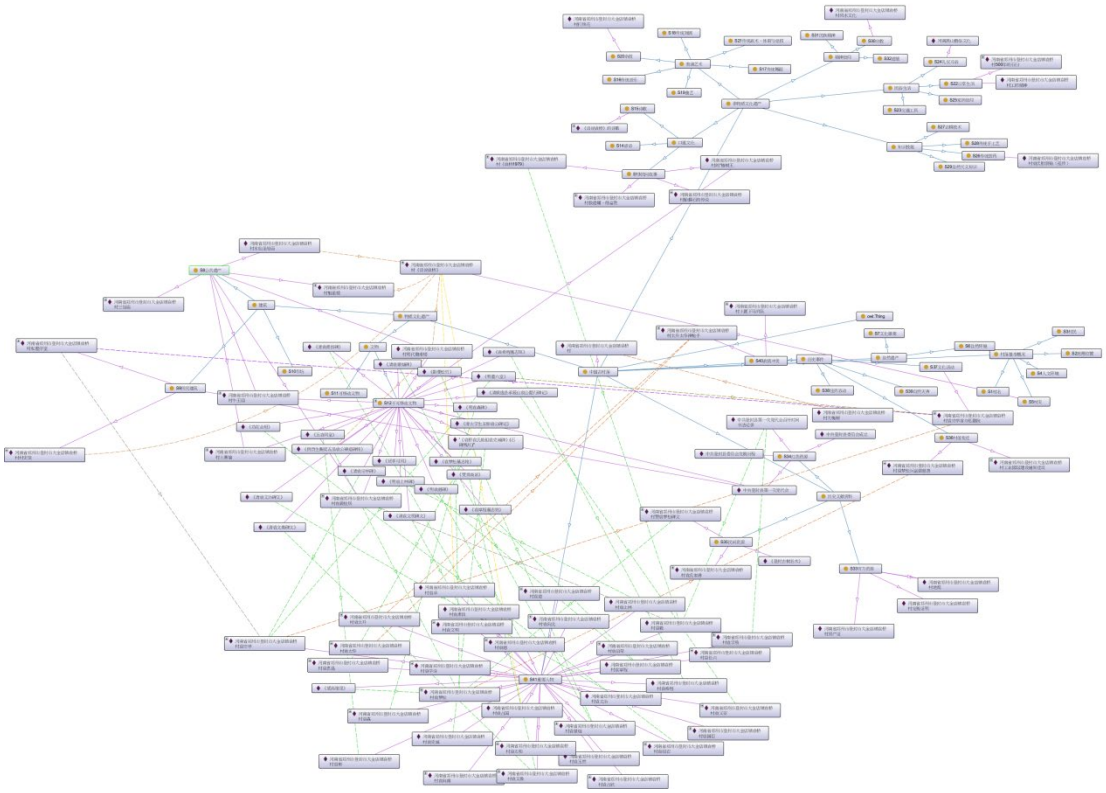


Figure 4. Semantic structures presented in the digital archives of Yuanqiao village

Youdian village, situated in Shanhuo Town, Yanshi District, Luoyang City, Henan province, is notable for its distinctive sunken courtyards and well-preserved immovable cultural relics such as stone tablets, plaques, ancestral halls, and temples. The Hua family, with its 22-generation genealogy, is central to the village's history. The village chronicle '*Hua family genealogy*' documents the village's history, biographies, public heritage, and the unique historical narratives of its stone tablets in great detail. In the resource framework diagram, '*Youdian village chronicle*' forms the core, supported by immovable relics and public heritage. The diagram illustrates the village's historical development, traces of its prominent figures, and the literary corpus centered on its numerous stone tablets, presenting a vivid and multidimensional portrayal of Youdian Village.

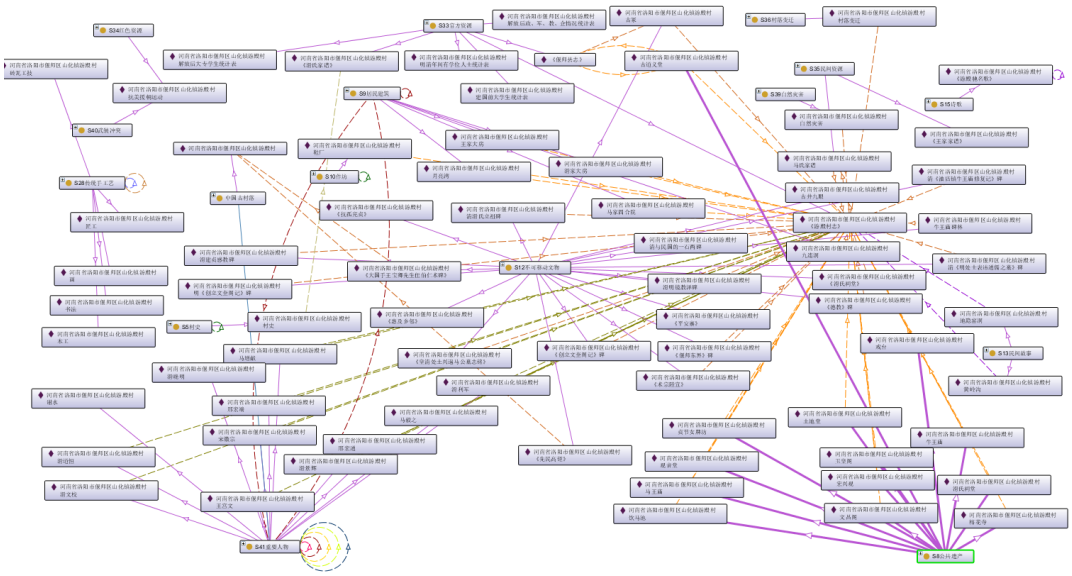


Figure 5. Semantic structures presented in the digital archives of Youdian village

Dawangmiao village, located in Tantou town, Luanchuan county, Luoyang city, Henan province, was once a campus of Henan university. The site still preserves the former women's dormitory of the university and holds rich red cultural resources, although some public heritage has suffered damage. The village's archives focus on its wartime history and Henan university's educational legacy. In the resource framework diagram, prominent figures are mainly educators and teachers, while most immovable cultural relics are associated with Henan university, including the women's dormitory and teaching tools. Folk stories center on red narratives of resistance against foreign invaders during the war. These elements collectively highlight Dawangmiao village as a red cultural site rooted in patriotism and national sentiment.

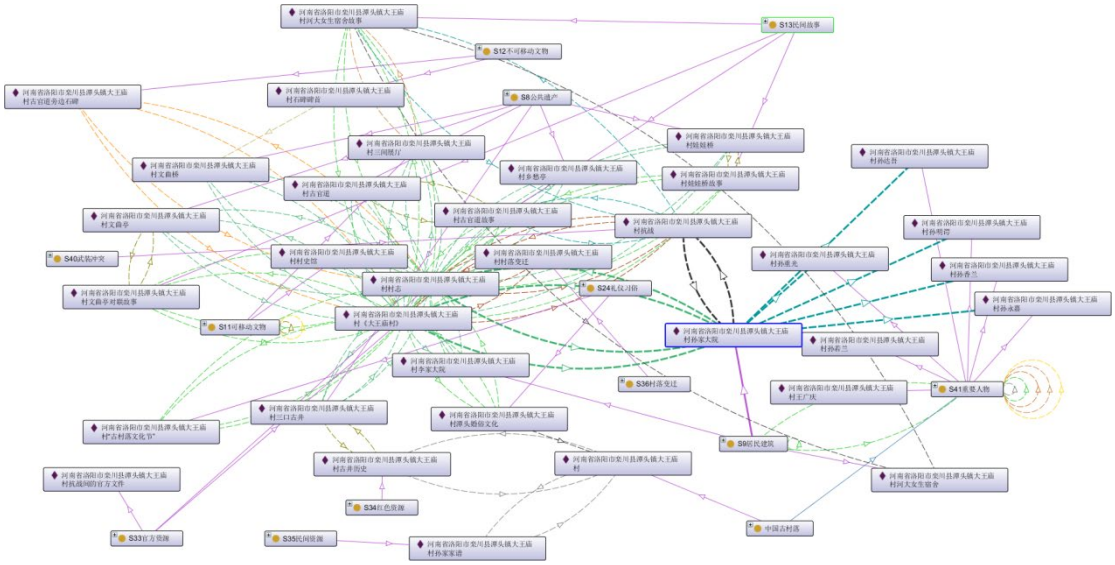


Figure 6. Semantic structures presented in the digital archives of Dawangmiao village

Future presentation of digital archives of ancient villages for cultural narratives

Based on the extraction of core cultural elements from the archival standards of Chinese ancient villages, this study employs internationally recognized semantic frameworks and cultural analysis models to construct a semantic analysis framework for digital archives of ancient villages. Using the digital archival resources of four villages in Henan Province as a case study, it explores a pathway for cultural knowledge discovery rooted in the digitization, semantic annotation, and presentation of ancient village archives.

However, cultural knowledge discovery is only the initial step. The challenge lies in effectively presenting this knowledge within digital spaces. Research teams must focus on designing immersive and interactive narrative scenarios grounded in the spatial storytelling of ancient villages. Future directions may include developing an interactive knowledge-based Q&A system that integrates online and offline experiences, educational competitions, and personalized digital guides for ancient villages. By leveraging technologies such as VR and AR, immersive exhibition halls for ancient villages could be created, along with interactive cultural games aimed at decoding historical elements. As digital technologies continue to advance, this framework has the potential to be further optimized, making significant contributions to the preservation and promotion of ancient village cultural heritage.

Acknowledgements

This paper would like to thank Professor Feng Huiling, a national first-class professor, and the digital memory team of Renmin University of China led by her. Since 2016, the digital memory team has continued to carry out research on the archiving and digital archival management of ancient villages in Zhejiang Province, laying a good research foundation for this paper. In addition, it is appreciated that Henan Provincial Archives, Zhengzhou Municipal Archives, and other institutions provided sufficient research guarantee and data support for the empirical study of Henan four villages.

About the author(s)

Tianjiao Qi is associate Professor in School of Information Resource Management and one of researchers in Research Center of Digital Humanities of Renmin University of China. She received her Ph.D. from Renmin university of China, and her research interests are semantic organization of digital archives, digital cultural heritage, and digital memory. She can be contacted at qtjjoy@163.com.

Yiduo Wang is a bachelor student in the School of Information Resource Management at Renmin University of China. She can be contacted at 2022200983@ruc.edu.cn.

Meiman Wong is a bachelor student in the School of Information Resource Management at Renmin University of China. She can be contacted at wongmeiman@qq.com.

Tsamchu Basang is a bachelor student in the School of Information Resource Management at Renmin University of China. She can be contacted at bsc2022202829@163.com.

Linglu Li is a bachelor student in the School of Information Resource Management at Renmin University of China. She can be contacted at 18888525801@163.com.

References

An H.Q., (2017). A preliminary study of traditional village archives management work. Beijing Archives,2017(05):29-30. https://doi.org/10.3969/j.issn.1002-1051.2017.05.009

CIDOC. what is the CIDOC CRM? https://cidoc-crm.org/

Ciotti, F., (2016). Toward a Formal Ontology for Narrative[J]. Matlit Revista do Programa de Doutoramento em Materialidades da Literatura ,20164(1):29-44. https://doi.org/10.14195/2182-8830_4-1_2

Digital Museum of Chinese Traditional Villages. Chinese traditional villages digital museum village building filling instructions. https://wj.dmctv.cn/index/中国传统村落数字博物馆村落建馆填报说明.pdf

FAN,Q.,TAN,G.X.,SUN,C.M.,(2023) Research on video resource description and knowledge organization of intangible cultural heritage. Journal of Central China Normal University(Natural Sciences) ,2023,57(03):469-482. https://doi.org/10.19603/j.cnki.1000-1190.2023.03.017

HU,Y., CHEN,S.,CAO,W., CAO,C,Z., (2014) The Concept and Cultural Connotation of Traditional Villages. Urban Development Studies,2014,21(01):10-13. https://doi.org/10.3969/j.issn.1006-3862.2014.01.021

Luo Z.Y.,(2019).Unified implementation of filing and listing protection for ancient towns, vill ages, and residential buildings throughout the province.Headline. https://www.toutiao.com/article/6711169247831130638/?upstream_biz=doubao&source=m_redirect&wid=1736261058380

He,S.Y., (2017). Guarding the Rvyal Memory: Study on the Archiving of Traditional Village.Archives Science Study ,2017(05):49-53. https://doi.org/10.16065/j.cnki.issn1002-1620.2017.05.009

Liu,H.Z.,Bao,H.,Yu,J.H., (2006). A Semantic Web Architecture of Virtual Museum Based on CIDOC CRM. Application Research of Computers ,2006,(04):50-53. https://doi.org/10.3969/j.issn.1001-3695.2006.04.017

Man,Y.,Wang,P., (2017). Present Situation and Strategy Research of Traditional Village Archiv aI Protection. Archives & Construction,2017(07):12-15. https://doi.org/10.3969/j.issn.1003-709 8.2017.07.004

National Archives Administration of China. (2021). Research on the Digital Memory Construction of "Ancient Villages in Taizhou". National Archives Administration of China. https://www.saac.gov.cn/daj/kjcgtg/202102/1a43b031ef3d4ef08bb12e3c66fb5a5f/files/0bb22f 3af6844ae7b5fd9f815154ec66.pdf

Raimond, Y., Abdallah, S., (2007). The Event Ontology.https://motools.sourceforge.net/event /event.html

Tang, Q., (2023) Research on archival protection and governance of traditional village cultural heritage based on living heritage. China Archives,2023, (10):66-68. https://kns.cnki.net/kcms2/article/abstract?v=YK1b_MowDz8P0ikNU4izkQyVvLor_QCrpq7m uznj7QSDXAN6j9mtqsCE_oqrq3UB_q5yshaFdyoWSVHVHFaJQ6RdPQLPQq-OGZFIIxBd-Q6Eg2jOiJjijxlfVStJ_HGunHqA8nSk_zkN2ez9Fd036nH9Q3Re6UvzNQQTGjfWxl4=&uniplatform= NZKPT

Traditional Chinese Village Digital Museum.Traditional Village Weekly Information Essence (November 2, 2024- November 8, 2024).https://www.dmctv.cn/zxShow.aspx?id=3454

Traditional Villages Network. Traditional village archiving survey (text) filing form. http://www.chuantongcunluo.com/index.php/Home/Lddc/bgxz_wz.html

Traditional Villages Network. Traditional village archiving survey (picture) archive table. http://www.chuantongcunluo.com/index.php/Home/Lddc/bgxz_tupian.html

Xinhuashe, (2023) First Watch|Traditional village protection, general secretary emphasizes. https://news.cyol.com/gb/articles/2023-04/28/content_4wlPRVIWlN.html

XU, X.Y., LIU, X.X., (2017) Interpretation of Connotation and Differences between Ancient Villages Archives and Rural Archives.Archives Science Study.https://doi.org/10.16065/j.cnki.issn1002-1620.2017.04.008

Zheng, J.X., Luo, Y.G., Han, Y.H., (2023). Research on the Document System of Multi-Subjec t Collaborative Management of Traditional Village Cultural Heritage. Comparative Study of Cultural Innovation,2023,7(34):194-198. https://kns.cnki.net/kcms2/article/abstract?v=YK1b_ MowDz9k0ew_GX_xiEqk27ItNsGTR9RBq4zR7hvQ8yll0zfKrMtERm2_v9Smlqb0FE6RrvQwHW--I S4joZ5qyN9be17y9dq3dW9oIit-C7Yzd2Ad1a-Zpg4BZ1_xQk4hUuqv3Jy6j75HqQaqylzwcFjRnUuB_C Dy26k_Ulo=&uniplatform=NZKPT

Information Research - Vol. 30 No. iConf (2025)

A cultural heritage data curation framework for knowledge discovery

Ruijie He and Xiaoguang Wang
DOI: https://doi.org/10.47989/ir30iConf47323

Abstract

Introduction. Cultural heritage data resources (CHDR) serve as an important foundation for the revitalization of cultural heritage. By systematically investigating the construction process of CHDR, it helps to further clarify the resource construction process and promote the construction of CHDR oriented towards openness, collaboration, and intelligence.

Method. By means of literature review, case analysis, open coding, and other methods, 105 documents and 75 projects were analysed.

Analysis. On the basis of d-KISTI Model, a series of technical methods of CHDC are summarized from the literature and used as coding reference for further project analysis. Through the analysis of project cases, the processing and construction process of cultural heritage data resources are summarized and sorted out.

Results. A cultural heritage data curation (CHDC) framework for knowledge discovery has been proposed. There is still a broad scope for improvement in aspects such as structured expression, multi-sensory interactivity, sharing mechanisms, and copyright protection in the construction process of CHDR.

Conclusion(s). This paper focuses on the development needs and reality of resource construction in the field of cultural heritage from the perspective of knowledge discovery, as well as emphasizes the importance of collaborative opening and embracing technology in this process.

Introduction

With the proliferation of data volume and the advancement of data processing technology, data has become a core resource for research in various fields, driving the exploration and discovery of new knowledge and new laws. Against the background of the initial success of cultural heritage resources digitization, further datafication construction has become a new impetus for data-intensive academic research and an important foundation for cultural heritage to move towards intelligence. By understanding the current status of cultural heritage datafication and sorting out the process and methods of CHDR construction, we can further clarify the needs and deficiencies of data construction, promote data reuse and discovery through data resources construction, and help the field of humanities adapt to the development of the data era.

Considering that cultural heritage resources will be preserved for a long time after digitization and datafication as important basic resources for subsequent use, this paper will combine the data life cycle and data curation related theories and methods, summarize the cultural heritage data curation methodology process through literature research, analyse the related projects from data collection to data preservation and utilization, and examine the current situation of cultural heritage resources through the perspective of knowledge discovery. We will examine the current status of CHDR construction from the perspective of knowledge discovery, construct a framework for CHDC oriented to knowledge discovery, summarize the current situation and shortcomings, and provide a reference direction for optimization and improvement of CHDR construction and cultural knowledge discovery.

Literature review

Lifecycle, data lifecycle and cultural heritage data lifecycle

Life cycle refers to the entire process of an object from its creation to its demise. '*Once something has the three important attributes of continuity, irreversibility, and iteration, it can be considered to be examined from a life cycle perspective*' (Ma & Wang, 2010). The concept of the data life cycle, which focuses on the entire process of data from its creation to its destruction, emphasizing the characteristics of data and its applications at different stages and aiming to ensure its sustainability, has played an important role in bridging the gap between data management and research.

Resources building in the field of cultural heritage includes resource collection, processing, transmission, use and storage. Typically, the links are continuously iterative and cannot be skipped, which is in line with the scope of application of the life cycle. It is worth noting that most cultural heritage resources do not exist in extinction or destruction. This stage is instead replaced by the permanent preservation of cultural heritage resources. Nowadays, the construction of data resources for cultural heritage has become an important foundation for the long-term development of digital humanities. Data curation based on data lifecycle is more in line with the concerns of preservation, maintenance, and care in this field. This paper defines the lifecycle of CHDR as,

> *the entire process from the creation to the permanent preservation of CHDR, encompassing stages such as collection, processing, storage, and utilization, with the ultimate goal of achieving value-added and reusable data, characterized by its permanence and iterability.*

Data curation and cultural heritage data curation

Data curation initially emerged as a response to challenges related to long-term data preservation, scientific utilization, and effective value addition. It focuses on comprehensive management of data with a clear lifecycle, involving planning and management at the point of data creation. Throughout the entire lifecycle, active management and evaluation are employed to maintain and

enhance the resource for both current and future use, thereby achieving value addition (Feng & Richards, 2018; Lee et al., 2020). In recent years, as the domains of data resources have diversified, governmental data (Gao et al., 2020) and humanities data (Laksmi et al., 2024) have gradually become noteworthy digital curation objects.

In the field of digital humanities, '*effective management and preservation of cultural heritage data contribute to the full realization of the value of cultural heritage resources*' (Laksmi et al., 2024). '*In general, data curation and digital humanities have similarities in terms of future goals and practical paths*' (Poole, 2017). '*The foundational construction and sustainable utilization of data resources are important issues of concern for both*' (Gao & He, 2023).

Knowledge discovery

'*Knowledge discovery is defined as the entire process of discovering useful knowledge from data*' (Fayyad et al., 1996). '*Earlier studies focused on data identification and data utilization*' (Guo, 2022). In recent years, as a goal-oriented theoretical perspective, knowledge discovery has been used in the construction and optimization of technical system platforms and digital service systems (Hu et al., 2022; Zhao et al., 2022), and combined with technical means such as knowledge mapping, artificial intelligence, machine learning (Han & Sun, 2021), which is regarded as an important value-added process of data (Hao & Gu, 2023). Knowledge discovery can uncover the value and application potential embedded within CHD, leading to the effective enhancement of CHD resources, and promoting in-depth applications of CHD across various fields.

Therefore, this paper integrates data lifecycle theory and data curation concepts to explore the full lifecycle management and value realization of cultural heritage data. By analysing and researching the stages of data generation, collection, processing, storage, and utilization, and combining the significant value-adding role of knowledge discovery for cultural data, a CHDC framework oriented towards knowledge discovery is constructed. Furthermore, by delving into the current practical situation of CHDR construction, this paper thoroughly discusses the management, sustainable value addition, and reusability of CHDR. It provides theoretical support and practical guidance for building a sustainable ecosystem for digital humanities and cultural heritage in the digital space.

Methodology

Foundation model

As a classification method for data curation-related operations, the d-KISTI Data curation Life cycle Model (d-KISTI Model) (Rhee, 2024) draws on the DCC model. A series of operations in the process of data curation are divided into sequential operations (such as creation and collection, evaluation and selection, extraction, and preservation, etc.), accidental operations (such as re-evaluation, disposal, migration, etc.), and life-cycle operations (such as referring to the core and key parts of data curation work, respectively). And the various stakeholders and their actions, in addition to the data itself, that may influence the data curation process. This model has been widely consulted in the process of design and development and has been verified in practice in the environment of digital library. It encourages resource users to examine the preservation and utilization of data resources from the perspective of methodology, and conforms to the goal of opening, protection, and utilization of cultural heritage resources construction. Therefore, this paper chooses this model as a basic reference, examines the methods and processes of cultural heritage data resources construction, and explores the current situation of data resources construction in combination with project practice.

Document analysis

As a type of literature review coexisting with narrative review, systematic review is able to collect relevant research information in a more comprehensive and specific way,

> *analyse and evaluate the literature that meets the criteria through a series of principles and methods* (Qiu, 2010).

'*Systematic review has the characteristics of problem orientation, screening criteria, comprehensive search, and effective quality*' (Aromataris & Pearson, 2014). This method helps researchers to find the most comprehensive 'evidence' possible in relation to the research question and has been widely used in research on topics such as information behaviour (Wang, 2018) and digital services (Tu & Liu, 2023).

> *The systematic review method can be roughly divided into the basic steps of problem identification, literature search, critical assessment, extraction and integration, interpretation, and analysis* (Aromataris & Pearson, 2014).

It can also be combined with looping, backtracking and other operations to enhance the comprehensiveness of the data sources, and to continuously revise and improve the research analysis and conclusions. This paper adopts the basic idea of systematic review and collects and refines the methods and techniques in the process of building CHDR based on the d-KISTI model.

Retrieval strategy

Digital resources are foundational to data resources in the CDH resource construction process, making digitization essential. This paper includes '*digital resource*' and '*data resource*' in its search scope, employing terms like '*cultural heritage*,' '*digital resources*,' '*data resources*,' and their translations, along with Boolean operators and wildcards for an effective search strategy, focusing on the '*subject*' field to enhance search relevance and coverage. Finally constructing the basic search formula as follows: Chinese literature search formula: SU%= '文化遗产' AND SU%=('数字资源' + '数据资源' + '资源建设'), English literature search formula: TS = '*cultural heritage*' AND TS = ('*digital resource**' OR '*data resource**' OR '*resource* construction*'). We selected China national knowledge infrastructure (CNKI) database and web of science (WoS) core collection as the search source. We also tracked the citations and references that met the requirements, supplemented with Google scholar to obtain the citations and references that were not in the selected databases without limiting the time of publication, and used the snowball tracking strategy to minimize the omissions due to the incompleteness of the search terms.

Assessment and screening

According to the research objectives and research questions of this paper, the reference standards for literature evaluation and screening are determined as follows: the construction process of CHDR is taken as the research content, including the construction of methodology or theoretical system, the research related to the practice of project platform construction, and the type of literature is limited to the journal papers and conference papers in both Chinese and English. First of all, using the Chinese literature search formula to search CNKI database, allowing synonym expansion, we got 1059 pieces of literature, from which 759 journal papers and 40 conference papers were screened out, and by browsing the titles, keywords and abstracts, we initially got 75 journal papers and 2 conference papers, and supplemented 5 pieces of literature after tracking the cited literature and references to the above 77 papers, which amounted to a total of 82 pieces of Chinese The total number of Chinese papers was 82. Then we used the English literature search to search the WoS core collection, and obtained 118 papers, including 76 journal papers and 42 conference papers, screened 10 journal papers and 7 conference papers, and supplemented 6 cited papers and references, totalling 23 English papers.

Extract and integrate

Read the full text of the 105 selected documents, draw on the methods and ideas of grounded theory, use the d-KISTI model as the first-level coding of the reference frame for document information extraction, and extract and collect the second-level coding corresponding to each

operation stage through literature reading. Due to space limitations, some open coding results are shown in Table 1.

One-level coding	Secondary coding	Partial source data
Conceptualise	Metadata	It is necessary to make a metadata description of the stele and rubbings to reveal the content features of the stele such as title, responsible person, summary, writing style, and material characteristics such as size and material.
		The main goal of this research is to develop a model to aggregate diverse CHI resources on the Web based on One-to-One Principle of Metadata. The One-to-One Principle helps to distinguish digital copies and their source related to cultural heritage objects.
Create & collect	Crowdsource collaboration	Drawing on the concept of Internet co-construction and sharing, museums should build an open platform for the public and use the wisdom of the masses to crowdsource data to build digital humanities projects.
		The construction of intelligent data cannot be separated from group collaboration and social crowdsourcing. Build a crowdsourcing platform or online community for intelligent data collation, processing and organisation in the field, and the public participate in data crowdsourcing through social networks.

Table 1. Document coding results (part)

Analysis

Combining the results of literature reading and information extraction, the CHDC method process is summarised, as shown in Figure 1.

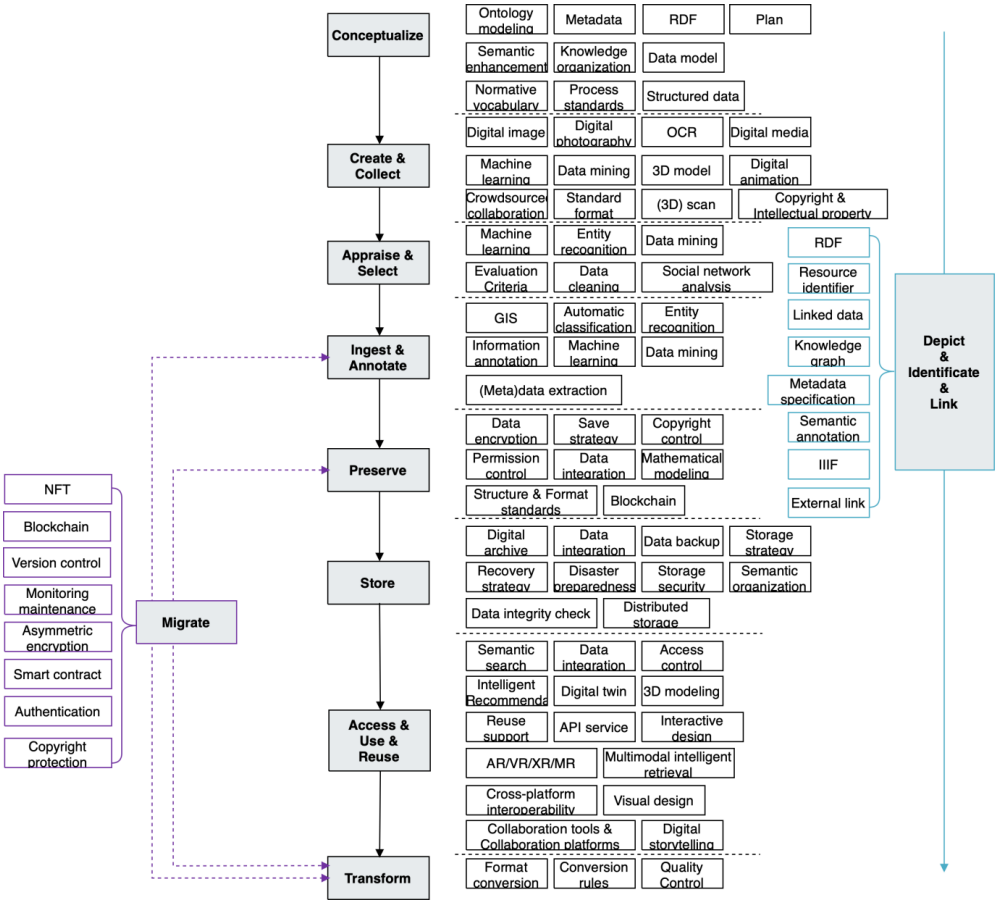


Figure 1. CHDC method and process

Sequential operations are represented by black boxes and connected in sequence by solid black arrows. The conceptualisation stage aims to apply knowledge organisation methods to construct framework standards, develop metadata specifications in combination with international standards create an ontology model, utilise the list of subject headings to carry out semantic extensions, and ultimately establish a systematic framework for describing the resources. In the specific process, the three stages of creation and collection, appraisal and selection, and ingestion and annotation are usually executed simultaneously. First, digitize physical objects through photography and scanning to form a cultural heritage database in a specific field. Then combined with machine learning to accomplish entity recognition, automatic categorisation, data extraction and other data mining and cleaning, as well as the semanticisation based on the extraction and annotation of resources. Only three papers mentioned the appraisal and selection, mostly with the degree of topic relevance as the default measure. It is noteworthy that crowdsourcing collaboration has become one of the effective modes of CHDR collection and processing, but the issue of copyright and intellectual property rights in the process of data collection has also come to the forefront. In order to realize the long-term preservation of CHDR, it is also necessary to consider appropriate preservation methods and storage strategies, and to realise data integration and reuse with as low a capacity as possible, low risk, high efficiency, multimodality, and strong correlation. Data backup, integration and security are essential requirements for cultural heritage resources management, combined with integrity checks to ensure data validity and availability, resulting in standardised digital archives. Semantic organisation and correlation of data are key to storage design, while blockchain encryption methods facilitate control and tracking of copyright and ownership. In addition, meeting the needs of different types of users to participate in the exploration and discovery process of CHDR through multi-dimensional and multi-sensory

experience interactions is a key link in realizing the value of datafication and an important stage in realizing the intelligentization of CHDR and the provision of intelligent services. In addition, multi-dimensional experiential interactions to satisfy users exploring CHDR are key to datafication value realization and intelligent service delivery. Interactive design, visualization design, virtual reality and its extension technology are widely used in the construction, display and utilization of various data resources, intelligent semantic search and recommendation based on multimodal data integration, and cross-platform interoperability design have been paid attention to, and the development and establishment of collaborative tools and cooperation platforms have become the key infrastructure for data openness and reuse. It is also possible to enhance the possibility of data access and reuse by transforming changes in the external form of data such as format, which is common in the transformation of metadata format and resource modality, but not much has been mentioned so far.

Occasional operations are located on the left side of the framework diagram and are connected to the possible stages of the operation by dashed arrows. Resource selection is mostly done at the '*assessment and selection*' stage and emphasizes attention to the flow of copyright and ownership during the migration process. The use of NFT, blockchain, smart contract, asymmetric encryption, and other methods to realize copyright traceability and identity authentication has been preliminarily explored and is considered for monitoring of CHDR.

The full life cycle operations are located on the right side of the framework diagram, with long straight lines running vertically through the sequential and occasional operations. The stage of description, identification and linkage aims to realize accurate description, identification, and linkage of cultural heritage resources by actively linking internal and external resources through RDF, Linked Data, Knowledge Graph and other methods and technologies, and by using them as a reference to coordinate and plan each attribute and characteristic of the data. This is not only an effective way to expand the data utilization scenario, but also an important basis for enhancing the possibility of data discovery.

Project analysis

Sample selection and coding results

Considering the relevance of the project, the impact of the award and the accessibility of the data, we have chosen the 2022 China digital annual conference excellent project cases of humanities (CDH) and award-winning projects of digital humanities awards (DHA) are selected as the research objects. Due to the small number of projects in CDH2022, excellent projects of CDH2021 are also included in the research scope. After removing one duplicate project, a total of 75 digital humanities projects became our research objects.

Based on the above series of summaries of the operation process of the literature related to CHDC, we will further combine the bottom-up open coding method to collect, visit, read and analyse the literature materials (such as papers, news, etc.) and platform websites of 75 projects. First of all, we take the names of each stage in Figure 1 as the reference of first-level coding, and the specific technical methods in it as the second-level coding to form the initial coding scheme. Then based on this, the project data is investigated and analysed, and the coding scheme is constantly adjusted, supplemented, and improved in the process. The final coding scheme was obtained as shown in Table 2.

Stage Name	Concrete Action	Number of Items
Standards, guidelines & specifications Design	/	2
Initial conceptualization	Metadata & Vocabulary Reuse	5
	Knowledge organization	6
	Ontology modeling	0

Stage Name	Concrete Action	Number of Items
Data collection & Creation	Physical data collection	31
	Digital data collection	23
	Crowdsourcing collaboration	7
	Evaluation & Selection	7
	Storage	2
Conceptual improvement	Vocabulary building	6
	Metadata specification	13
	Data & Model structuring	12
Extraction & Annotation	Extraction	10
	Annotation	4
Identification & Linking	External linking	12
	Internal linking	7
Presentation & Display	Accessibility	29
	Visualization	28
	Interactivity	23
	Experientiality	3
	Reusability	30
	Linkability	1
	Collaborability	11
Others	Copyrights & IPR	7
	Data transformation	1

Table 2. Coding scheme

Findings

The design of standards, guidelines and specifications is the first step. The operational content mainly focuses on digitization operation, process specification and operation guidelines, reflecting the transitional characteristics of the integration and construction stage of resources in cultural heritage resources domain towards the integration and utilization stage. The initial conceptualization stage does not seek to be exhaustive, but rather to design the basic framework of the resource and improve the interoperability of the data through the reference and reuse. The operations are mainly focused on the reuse of metadata sets and vocabularies and the initial establishment of the research project's knowledge organization methodology. The data collection and creation stage in this field has not only seen the traditional researcher-led model, but also the enhancement of the breadth and diversity of cultural heritage data through the power of the public. Most projects choose to digitize physical materials directly and use existing resources through correlation or extraction, avoiding the waste of costs associated with repeated collection of similar information.

Data appraisal and selection are the key to ensuring the quality of massive datasets, with key metrics including copyright, accuracy, and uniqueness. Copyright is concerned with ensuring legitimacy, accuracy assessment ensures data quality, and uniqueness requires enhancement of the character and value of CHDR. In terms of data storage, cultural heritage data projects are concerned about the association between data and databases, for example, by matching metadata tables with encoding index tables to help facilitate join operations.

The refinement of the conceptualization stage involves extending the initial word list framework and constructing specialized metadata specifications in conjunction with standardized word lists. Given the diversity and complexity of cultural heritage data, researchers need to customize the ontology to accommodate domain specificity and integrate data from different sources to achieve

heterogeneous data mapping and integration. The extraction and annotation stage enriches the data content and improves the information density through key information refinement and data annotation, including text entity relationship extraction, image recognition classification, and enhances the data semantics by combining with the open annotation model, while actively adopting the annotation method combining manual and machine learning. The identification and linking stage build a complex semantic network through internal and external data association for effective resource identification and linking. External linking utilizes databases such as entity names or Wikidata or interoperability frameworks, while internal associations use knowledge graphs and keywords to facilitate multidimensional analysis and data exploration.

The task of the external presentation and display stage is to present processed cultural heritage data to the user in a rich and understandable way and providing interactive channels to meet diversified needs. This stage consists of seven aspects: accessibility emphasizes open access to the resource, visualization enhances the visual experience through static and dynamic means, interactivity allows users to explore the resource and have a personalized experience, experiential creates immersive environments through digital technologies, reusability considers the re-creation and sharing of the resource, linkability emphasizes the overall openness of resources, and collaborability supports interaction and cooperation among users. Although some projects achieve these aspects, there is still room for improvement in the areas of open licensing, API usage, and user community building. In addition, some projects have focused on copyright and intellectual property protection and data transformation.

In general, the operations involved in CHDR building projects are adapted to the specific needs of the project, but usually include the development of digitization specifications, technical collection, and conversion of resources, ensuring data copyright and quality, refining conceptualization and structured representations, as well as enhancing linkages and multi-format presentations of the resources, and ultimately, the sharing and use of the resources.

A CHDC framework for knowledge discovery

Based on the literature review of life cycle and data curation, we carried out literature reading and collating on the theme of CHDC, obtained a series of specific technical methods, and summarized the main methodological framework of CHDC based on the investigation and analysis of 75 project cases. Since the focus of CHDC is to satisfy the knowledge mining and utilization of cultural heritage resources, we believe that the concept of knowledge discovery can be used to emphasize the goal of this process, and a framework for CHDC oriented to knowledge discovery is proposed, as shown in Figure 2.

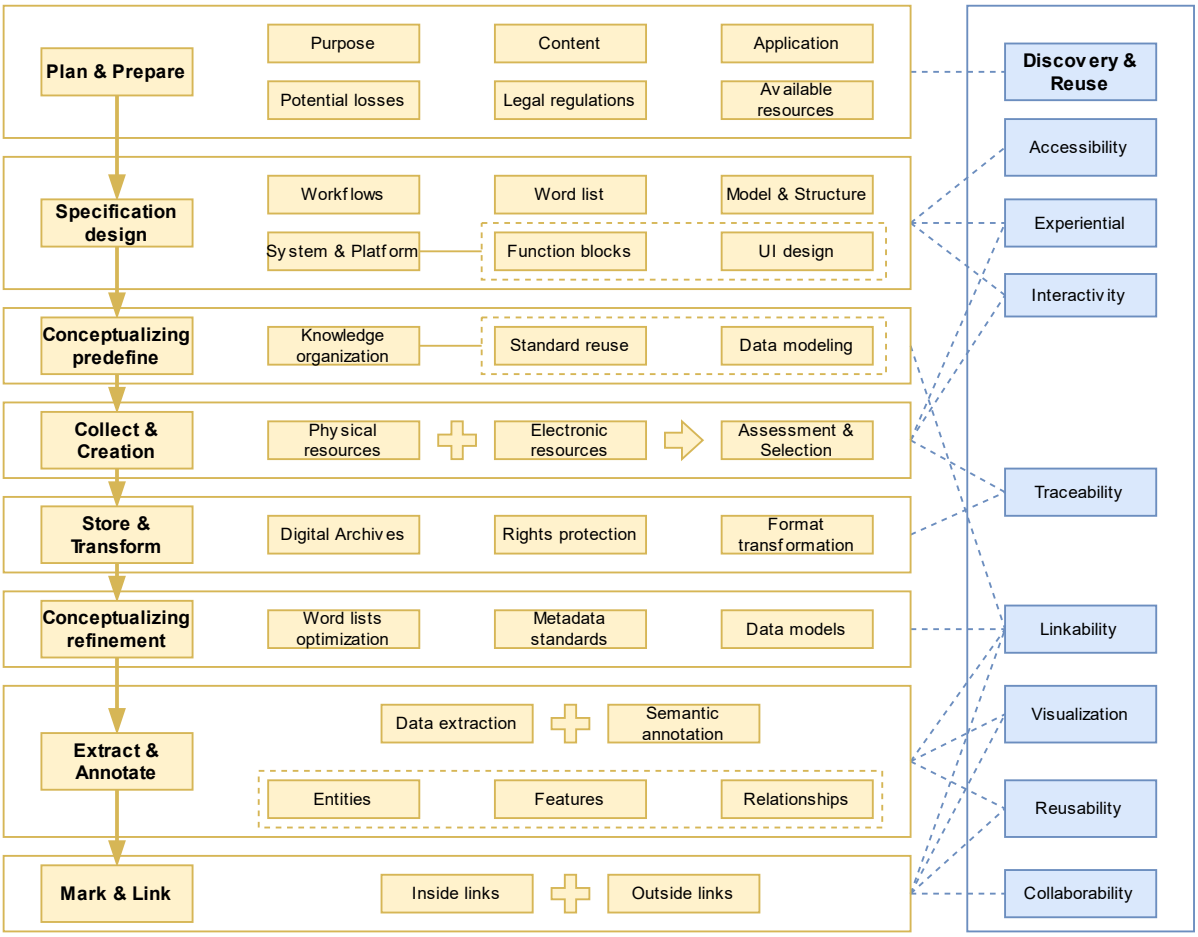


Figure 2. A CHDC framework for knowledge discovery

Explanation of the content of the framework

In the stage of planning and preparation, the project team needs to predefine the objectives, content, and use of cultural heritage resources, assess potential damages, clarify legal norms, and research existing resources. The aim is to provide a framework proposal of overall objectives and construction models for knowledge discovery and reuse. The specification design stage needs to develop specifications for data collection and processing, research and organize metadata standards and ontology models. Start early in the project and iterate with implementation to ensure that the dataset or system platform helps users to accurately understand and utilize the data. The conceptualizing predefinition stage requires the formation of metadata sets, word lists, and ontology models suitable for the project based on research and standards, to enhance the discoverability and availability of data, and to prepare for the open sharing of data. The stage of data collection and creation aims to collect and screen physical and electronic resources through technological means to create a digital archive and lay the foundation for a data resource for knowledge discovery potential. Attention also needs to be paid to copyright and data traceability. The stage of storage and transformation focuses on the traceability of copyrights and intellectual property rights, providing tools, rules and mechanisms for format transformation and safeguarding the project's own copyrights and intellectual property rights. The stage of conceptualizing refinement focuses on refining and perfecting word lists, metadata standards, and data models to form a complete and consistent data model and working to support data linking and retrieval. The stage of extraction and annotation involves extracting and annotating data entities, features, and relationships, integrating different data sources, and enhancing the semantic richness of the data. At the same time, it is necessary to consider the possibility of displaying the discovery results

through visualization methods to extend the data application scenarios. The stage of marking and association focuses on realizing the links between inside and outside resources, facilitating the integration of different modal resources, and enhancing comprehensibility and discoverability. It is also necessary to realize semantic enrichment through methods such as knowledge mapping to enhance the explorability of resources as much as possible. The stage of discovery and reuse is the core part of the later stages of building CHDR, focusing on user experience and engagement. And the discovery and reuse of data value is realized through resources, tools, and methods.

Characterization of the framework

Examining CHDR building through the data curation lens

Based on the analysis of the specificity of the life cycle of CHDR, this paper investigates the methodological process of CHDR construction and establishes a CHDC framework accordingly with reference to the d-KISTI model. On the one hand, '*traditional data lifecycle theory and data curation methods focus on data in a broad sense or big data, mainly focusing on research data and scientific data*' (Yoon et al., 2022), and less on the exploration and construction of cultural heritage data. On the other hand, the traditional construction of CHDR lacks a unified methodology and process specification, and the introduction of the data curation method can provide a more complete process of data collection, organization, analysis, display and utilization. On the other hand, the traditional CHDR construction lacks unified methodological process specification, the introduction of data curation method can provide a more complete process guidance from data collection, collation, analysis to display and utilization, which can help to systematize and standardize the process of data resource construction in this field.

Explaining the process of building CHDR around knowledge discovery

With the help of knowledge discovery, the potential value of CHDR can be more fully realized. Combined with the demand and realization of knowledge discovery, this paper proposes to integrate knowledge discovery into each stage of CHDR construction, so as to realize the continuity and sustainability of the whole process from data collection to knowledge discovery. At the early stage of the construction process, it emphasizes the predetermination and target determination of knowledge discovery, and provides support for knowledge discovery through specification design, data collection and processing, and conceptualization construction, so as to give users more ability to discover and explore the resources.

Emphasizing copyright and intellectual property rights for CHDR

'*Data curation explicitly requires the specification of rights and accountability*' (Brink, 2017) and the adoption of appropriate technical means to protect the security and legitimacy of data. In this paper, the active attention and protection of copyright and intellectual property rights are explicitly proposed in the framework of CHDC, and how to incentivize the creation and reuse of cultural heritage data and safeguard the legitimate rights and interests of resource owners or providers still need to be actively explored in conjunction with data encryption and transmission technologies.

Combining literature research and project platform practice

In the process of framework construction, this paper fully combines two means of literature research and project research. Through extensive literature research, in-depth understanding of CHDR construction related theories and methods and technologies, to provide a basis for reference for the framework construction; at the same time, combined with the specific practice of the project platform to verify and optimize the framework, to improve the relevance and applicability of the framework.

Future works

Actively promote universal structured expression design

Standardizing structured representations of CHDR by means of feature classification, metadata description, and ontology modelling facilitates knowledge storage, representation, and retrieval, and is essential for resource protection and intelligent application. Despite the current diversity of structured representations in the cultural heritage domain and the lack of unified standards, many projects choose to reuse and extend existing standards in similar domains. While this approach solves the problem of generalizability, there are still challenges of semantic inconsistency and cross-domain compatibility. In addition, further research is needed on how to make the organization of resources comprehensible to users from different backgrounds in order to attract them to explore the deeper laws and knowledge of cultural heritage.

Wide range of technological approaches to hyper-realistic multi-sensory interactivity

With the development of virtual simulation technology, the field of cultural heritage is shifting to scenario-based and experiential communication based on digital media, emphasizing multi-sensory and multi-angle immersive experiences. Panoramic image technology provides users with a virtual environment for close observation and hands-on experience, promoting the multidimensional presentation of cultural heritage information (Jiang & Zhang, 2023). However, current resource building projects mostly focus on data opening and display, lack interactive design from the user's perspective, and are mainly in the form of passive browsing. Cultural heritage projects can explore the intelligent generation and interaction of characters or scenes and utilize new digital media technologies to enhance the discoverability and exploration of knowledge, creating a more interesting and unique cultural experience.

Building resource synergy and open sharing mechanisms

There are deficiencies in the collaborative openness and sharing of CHDR, which are mainly manifested in the lack of platforms and standards for interoperability between projects, leading to the phenomenon of resource silos; an imperfect data ecology for user participation, limited channels for data access and reuse, and a lack of incentive mechanisms.

> *Realizing collaborative openness is a change in the way of knowledge production in the digital environment, involving the providers, deliverers, and users of resources, which requires the establishment of a support system*' (Wu et al., 2022).

The initial awareness of collaborative openness should be strengthened, translated into technical specifications and standards, drawing on excellent cases and designing an open platform, so that the concept of collaborative openness can be carried through the whole process of resource construction.

Optimize the means of copyright protection

'*The construction of culture heritage data resources involves original humanistic data, digitized materials of traditional literature and processed data*' (Ou, 2023), and the copyright issues are complex and risky. Uncertainty exists in the attribution of copyright of original materials, the legality of self-constructed materials, and the citation of third-party databases. The open standard of new-born resources is not uniform, the diversity of resource types leads to the complexity of the degree of authorization and openness, and the rules of open license agreements of different platforms are different. The definition of copyright in data reuse requires continuous exploration of legal policies and technical measures.

Conclusions

Digitization enhances the preservation, efficiency and value of cultural heritage resources and promotes the development of related construction projects. This paper systematically studies and

analyses the construction of digital CHDR, including standardized design, conceptualization, data collection, refinement, extraction and annotation, identification, and association, display and copyright, and proposes a framework for knowledge discovery. The research shows that although resource conceptualization and opening are more mature and interactive display has become mainstream, the diversity and complexity of cultural heritage resources requires customized solutions, while insufficient datafication platforms, limited application of virtual technologies, and copyright issues still constrain the development of large-scale collaborative knowledge discovery.

The digitization and datafication of CHDR is the key to their adaptive use, and in the era of artificial intelligence, sorting out process frameworks and standard specifications can help to improve the standardization of resource construction and promote data opening and sharing. The shortcomings of this paper lie in the possible bias in the understanding of the selected projects and the lack of expert support and empirical testing of the proposed framework. Future research will conduct empirical tests and practical applications, focus on the application of AI in cultural heritage resource construction, and collect expert feedback to conduct experiments from the user's perspective.

Acknowledgments

The authors would like to acknowledge the support of reviewers for their helpful comments. This research work has been supported by the Major Program of National Social Science Foundation of China (21&ZD334).

About the author(s)

Ruijie He, corresponding author, is a postgraduate student at the School of Information Management at Wuhan University, Wuhan, Hubei, China. Her research interests are digital humanities and data governance. Her contact address is ruijiehe_ch@163.com

Xiaoguang Wang is a professor at the School of Information Management at Wuhan University, Wuhan, Hubei, China. His contact address is wxguang@whu.edu.cn

References

Aromataris, E., & Pearson, A. (2014). The systematic review: an overview. AJN The American Journal of Nursing, 114(3), 53-58. https://doi.org/10.1097/01.NAJ.0000444496.24228.2c

Brink, J. A. (2017). Big data management, access, and protection. Journal of the American College of Radiology, 14(5), 579-580. https://doi.org/10.1016/j.jacr.2017.03.024

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37. https://doi.org/10.1609/aimag.v17i3.1230

Feng, Y., & Richards, L. (2018). A review of digital curation professional competencies: theory and current practices. Records Management Journal, 28(1), 62-78. https://doi.org/10.1108/RMJ-09-2016-0022

Gao, D., & He, L. (2023). Digital humanities researches from the perspective of data intelligence empowerment: data, technology, and applications. Library Tribune, 43(09), 107-119. https://doi.org/10.3969/j.issn.1002-1167.2023.09.014

Gao, G. Zhu, M., & Duan J. (2020). Research on government big data service standardization system based on data curation. E-Government, 216(12), 110-120. https://doi.org/10.16582/j.cnki.dzzw.2020.12.012

Guo, S. (2022). Research on library's knowledge service model of data-driven knowledge discovery. Journal of Library and Information Science, 7(05), 23-27. https://doi.org/10.3969/j.issn.1005-6033.2022.05.005

Han, T., & Sun, M. (2021). Intellectualization of scientific research and knowledge service: conceptualization, realization and opportunity. Information Studies: Theory & Application, 44(10), 41-49. https://doi.org/10.16353/j.cnki.1000-7490.2021.10.006

Hao, X., & Gu, X. (2023). Al reshapes the view of knowledge: knowledge creation and education development under the influence of data science. Chinese Journal of Distance Education, 43(05), 13-23. https://doi.org/10.13541/j.cnki.chinade.2023.05.002

Hu, Y., Li, X., & Zhu, X. (2022). The knowledge discovery of integrating subject words and citation: data optimization and content visualization. Journal of Intelligence, 41(10), 130-137+155. https://doi.org/10.3969/j.issn.1002-1965.2022.10.019

Jiang, N., & Zhang, J. (2023). The evolution, presentation, and management of cultural heritage information from the perspective of digital humanities. Journal of Tongji University (Social Science Edition), 34(06), 82-93. https://doi.org/10.3969/j.issn.1009-3060.2023.06.009

Laksmi, L., Suhendra, M. F., Shuhidan, S. M., & Umanto, U. (2024). The readiness to implement digital humanities data curation of four institutional repositories in Indonesia. Digital Library Perspectives, 40(1), 80-95. https://doi.org/10.1108/DLP-04-2023-0031

Lee, H., Yoon, S., & Park, Z. (2020). "SEMANTIC" in a digital curation model. Journal of Data and Information Science, 5(1), 81-92. https://doi.org/10.1108/10.2478/jdis-2020-0007

Ma, F., & Wang, J. (2010). A Literature Review of Studies on Information Lifecycle(I)——the Perspective of Value. Journal of the China Society for Scientific and Technical Information, (5), 939-947. https://doi.org/10.3772/j.issn.1000-0135.2010.05.024

Ou, Y.(2023). Risks and prevention strategies of data copyright in digital humanities application service. Journal of Library Science in China, 49(01), 118-128. https://doi.org/10.13530/j.cnki.jlis.2023008

Poole, A. H. (2017). "A greatly unexplored area": Digital curation and innovation in digital humanities. Journal of the Association for Information Science and Technology, 68(7), 1772-1781. https://doi.org/10.1002/asi.23743

Qiu, X. (2010). Systematic review: a more scientific and objective overview method. Documentation, Information & Knowledge, (01),15-19. https://doi.org/10.13366/j.dik.2010.01.016

Rhee, H. L. (2024). A new lifecycle model enabling optimal digital curation. Journal of librarianship and information science, 56(1), 241-266. https://doi.org/10.1177/09610006221125956

Tu, Z., & Liu, X. (2023). The contents and forms of digital scholarship services: a systematic review and comparative study. Library and Information Service, 67(08), 104-114. https://doi.org/10.13266/j.issn.0252-3116.2023.08.010

Wang, Y. (2018). User value of science and technology information: an integrative conceptual framework based on systematic literature review. Publishing Journal, 26(01), 82-89. https://doi.org/10.13363/j.publishingjournal.2018.01.018

Wijesundara, C., & Sugimoto, S. (2018). Metadata model for organizing digital archives of tangible and intangible cultural heritage and linking cultural heritage information in digital space. Library and Information Science Research E-Journal. https://doi.org/10.32655/LIBRES.2018.2.2

Wu, J., Shi, J., & Xu, J. (2022). Open data ecosystem for digital humanities: constituent elements and model frameworks. Library and Information Service, 66(22), 44-54. https://doi.org/10.13266/j.issn.0252-3116.2022.22.004

Yoon, A., Kim, J., & Donaldson, D. R. (2022). Big data curation framework: Curation actions and challenges. Journal of Information Science, 0(0). https://doi.org/10.13266/10.1177/01655515221133528

Zhao, Z., Liu, Y., Zhu, L., & Wu, X. (2022). Research on patterns and methods for knowledge construction and reuse in a complex information environment. Journal of the China Society for Scientific and Technical Information, 41(12), 1266-1279. https://doi.org/10.3772/j.issn.1000-0135.2022.12.005

Information Research - Vol. 30 No. iConf (2025)

Ideological presentations in official promotion of intangible cultural heritage (ICH) on short-form video platforms: a multimodal content analysis

Ruoxi Liu and Haoran Qiu

DOI: https://doi.org/10.47989/ir30iConf47317

Abstract

Introduction. This study investigates what ideological themes are embodied in the official promotion of intangible cultural heritage (ICH) on short-form video (SFV) platforms, and further examines how these ideological themes are presented through various content creation elements.

Method. Two trained researchers in information sciences conducted a multimodal content analysis of 532 SFVs downloaded from a nationally representative official Douyin (the Chinese version of TikTok) account promoting ICH.

Findings. This study identified four primary ideological themes embodied in the official promotion of ICH: (1) cultural democratization, (2) tradition revivalism, (3) patriotism, and (4) multiculturalism. Moreover, we revealed the distribution of content creation elements across SFVs reflecting various ideological themes.

Conclusion. In SFV platforms like Douyin, ICH promotion highlights grassroots-centered narratives and user-generated videography. Traditional and popular culture are integrated within this context. This fosters a relatively democratized discourse around ICH, creating a more inclusive and softened ideological expression.

Introduction

Intangible cultural heritage (ICH) refers to practices, expressions, knowledge, and skills recognized by people '*as part of their cultural heritage*' (UNESCO, 2022; Vecco, 2010), such as folklore, performing arts, craftsmanship, and languages, which can be transmitted to and recreated by future generations (Lenzerini, 2011). Nowadays, ICH has increasingly been recognized in terms of political, economic, and cultural values (Chmil et al., 2021; Cominelli & Greffe, 2013; Petronela, 2016; Tavares et al., 2021), highlighting its important role in demonstrating a nation's soft power (Rodrigues, 2023; Schreiber, 2017). Consequently, numerous official institutions have intensified efforts to promote and preserve ICH (Karampampas, 2021; Qiu et al., 2022; Xu et al., 2022).

Ideology can be defined as a belief system that reflects the interests, values, and goals of a particular group, class, or society (Converse, 2006; Lichtheim, 1965). Prior practices showed that the promotion of ICH can facilitate the construction and expression of political, cultural, and social ideologies. For example, the European Union has implemented a '*placing heritage*' strategy to utilize tangible and intangible cultural heritage in creating a collective European cultural value system across various countries (Lähdesmäki, 2016). Additionally, research suggested that ICH promotion is important for maintaining and strengthening multiethnicity and cultural diversity (Massing, 2018; M. N. Taylor, 2009; Vrdoljak, 2005).

For a long time, conventional media like film, photography, and print have been essential in ICH promotion (Sullivan, 2015). This kind of media content, especially those from official sources, often embodies dominant narratives and perspectives on the heritage that are sanctioned and promoted by authoritative bodies (Smith, 2006), which are usually used to advocate mainstream ideologies for governance (Macdonald, 2003). However, the advent of the digital era seemed to have redefined the ICH discourse and ideological expressions. Scholars suggested that digitization and social-mediazation have made the interpretation of ICH more polyphonic and less reliant on authoritative perspectives, incorporating more interpretations from the audience side (Kalay et al., 2007; J. Taylor & Gibson, 2017).

In recent years, the rise of short-form video (SFV) platforms such as TikTok has become a vital vehicle for the official promotion of ICH (Cao, 2022), which offers more appealing information formats and content creation (H. Wang et al., 2024). First, most SFVs adopt a vertical video format ranging from 30 to 60 seconds in length (Van Daele et al., 2024), which aligns better with smartphone users' habits and is well-suited for fragmented content dissemination (Ryan, 2018; Y. Wang, 2021). Second, compared to traditional media, the production of SFVs requires lower costs and has less stringent requirements for professionalism and aesthetics (Y. Wang, 2020). They focus more on micro-narratives and interaction, presenting a variety of engaging themes, expressions, and storytelling that appeal to the audiences (Y. Wang, 2021). Hence, SFVs may provide a more diverse, enriched, and even democratic ideological expression in the official promotion of ICH.

However, a notable research gap remains in how ideologies are embodied and presented in the official promotion of ICH on SFV platforms. More specifically, there is insufficient knowledge about the relationship among ideological presentations, ICH promotion, and content creation of official SFVs. Addressing this gap can foster a deeper understanding of the role of digitization and the social medialization of ICH in ideological construction. This can offer insights into the values and meanings of the official promotion of ICH from a cross-disciplinary and novel perspective, contributing to both social media and cultural heritage studies.

Hence, we proposed the following research questions (RQs) for this study:

> **RQ1)** What ideological themes are embodied in the official promotion of ICH on SFV platforms?
>
> **RQ2)** How are these ideological themes presented through various content creation elements of SFV?

Methodology

To address the proposed RQs, we conducted a multimodal content analysis of SFVs from a China's official account promoting ICH in Douyin (the Chinese version of TikTok). Choosing China and Douyin as the research entry points of this study is appropriate for two reasons. First, China is a multi-ethnic country with a rich variety of ICH resources. Currently, China has 1,557 national-level representative ICH items, 43 of which are listed on the UNESCO ICH list (UNESCO, 2024), ranking first in the world. Second, nearly all of China's ICHs are now promoted on the Douyin (Cao, 2022), attracting a large audience. A significant portion of this content is produced by official media outlets, such as national television stations, broadcasters, and state-affiliated organizations.

Data collection

We selected a representative official Douyin account promoting ICH for multimodal content analysis. As Gustafsson (2017) noted, focusing on a single case enables a richer and deeper exploration of a phenomenon compared to investigating multiple cases.

For representative account selection, we first searched on Douyin using the keywords '非物质文化遗产' (Chinese translation of '*Intangible Cultural Heritage*') and its abbreviation '非遗' (Chinese translation of '*ICH*'). Among the results, three criteria were applied for account filtering: 1) verified by official agency, 2) number of followers, and 3) number of likes received. Specifically, all unofficial accounts were first excluded. The remaining verified accounts were then ranked in descending order according to their number of followers. The top 10 accounts were further sorted in descending order based on their number of likes received. The account ranked first after this process was selected.

Finally, we identified the account '*Intangible Cultural Heritage is Coming*' (ICHC; Chinese name: '非遗来了'; ID: feiyitv). ICHC is an official Douyin account managed by *Radio Shanghai*, a provincial, state-owned media organization. This account was dedicated to promoting Chinese ICH, which has gained significant recognition and popularity nationwide. As of September 12, 2024, this account had the highest number of followers (1.483 million) and the highest number of likes received (7.729 million) among all official accounts promoting ICH in Douyin. A total of 532 SFVs published in this account were downloaded for multimodal content analysis.

Data analysis

We adopted Serafini and Reid's (2023) framework for multimodal content analysis. This framework (see Figure 1) includes 11 steps: (1) Recognizing an area of interest; (2) Developing initial research questions; (3) Constructing the data corpus; (4) Defining the object of study; (5) Developing initial categories; (6) Developing the analytical template; (7) Testing the analytical template; (8) Applying the analytical template to the data corpus; (9) Constructing potential themes; (10) Implications of the analysis; and (11) Dissemination of findings.

The rationale for selecting this framework lies in its extension of the traditional content analysis approach (e.g. Schreier, 2012) by integrating a multimodal theoretical perspective (e.g. Kress, 2009). This enables researchers to consider the interactions among various communication modalities, including video, image, text, and audio, which is particularly suited to the information characteristics of SFV content.

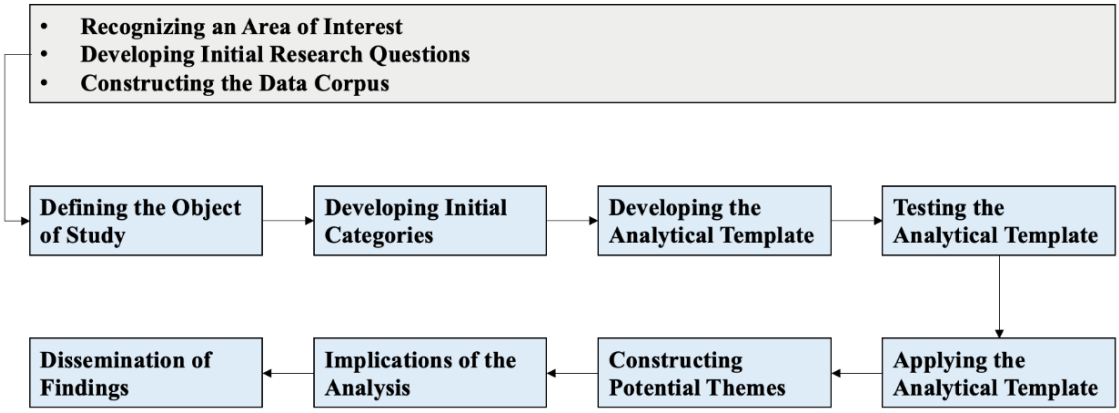


Figure 1. Serafini and Reid's (2023) approach to multimodal content analysis

To better address the proposed RQs, the analytical template (AT) for this study included two codebooks, each addressing RQ1 and RQ2, respectively. Both inductive and deductive coding approaches were used. To address RQ1, we applied Braun and Clarke's (2012) thematic analysis approach. This is a pure inductive coding approach. The codebook was developed based on the commonality of raw qualitative data. All generated codes were subsequently categorized, evaluated, and refined to construct appropriate themes. To address RQ2, we leveraged Kress's (2009) social semiotic theory of multimodality. This theory highlights that signs exist in various modes (e.g., visual, auditory, linguistic) and emphasizes the interplay among these modes in meaning construction. Drawing on this theoretical lens and considering the SFV's contexts, we proposed four primary categories: *narratives*, *videography*, *visual symbols*, and *music genres*. These categories served as a foundation for developing the elements within the codebook. Specifically, we further developed and confirmed corresponding elements (i.e., codes) of each category by coding a portion of the data. This process culminated in a finalized RQ2-specific codebook.

In the actual coding process, two trained researchers in information sciences systematically coded all SFVs (N=523). As shown in Figure 2, they first independently coded the first 100 SFVs and developed the initial AT. Subsequently, they discussed and resolved all inconsistencies and ambiguities of codebooks in AT. After that, they confirmed the AT. Guided by the AT, the two researchers worked collaboratively to code the remaining SFVs. Based on the coding results, they constructed a series of initial themes. In the last stage, codes with less than three occurrences were removed or merged. And coded themes were finally named, defined, refined, and confirmed. To ensure intercoder reliability, the coding process was repeated twice.

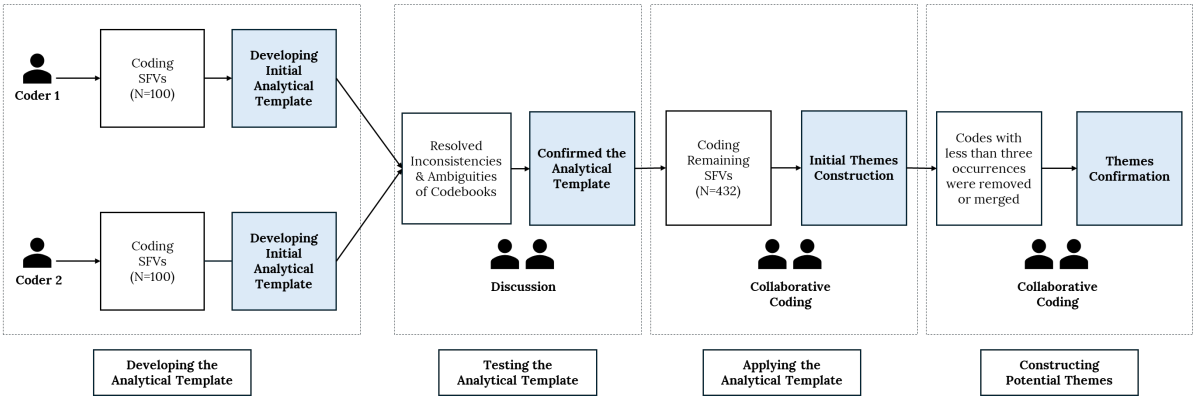


Figure 2. Workflow of the coding process

Results

For RQ 1, we identified four primary themes of ideological presentations in SFVs published by ICHC, including (1) Cultural democratization, (2) Tradition revivalism, (3) Patriotism, and (4) Multiculturalism. It is worth noting that the abovementioned themes were not mutually exclusive. Put differently, a SFV may present multiple ideological themes. Table 1 shows the distribution of themes in the RQ1-specific codebook of the final AT.

Themes	Frequency	Percentage of Total
Cultural Democratization	303	56.96%
Tradition Revivalism	198	37.22%
Patriotism	81	15.23%
Multiculturalism	65	12.22%
Total	532	100%

Table 1. Distribution of themes of ideological presentations

For RQ 2, we further explored how the abovementioned ideological themes were presented through the content creation of SFVs. Based on the four predefined categories, an RQ2-specific codebook with 11 elements (i.e. codes) of content creation was finally developed. Similarly, these elements were not deployed independently. In ICHC, most SFVs utilized a combination of multiple elements, reflecting the blending of various ideological themes presentations. Table 2 presents the statistics distribution of these elements across both the general set of SFVs and those featuring different ideological themes. Figure 3 visually shows these distributions using a Sankey diagram.

Categories	Elements	Cultural Democratization (N=303)	Tradition Revivalism (N=198)	Patriotism (N=81)	Multi-culturalism (N=65)	Total (N=532)
Narratives	Citizen Events	236 (77.9%)	67 (33.8%)	9 (11.1%)	8 (12.3%)	267 (50.2%)
	Ethnic Events	48 (15.8%)	127 (64.1%)	5 (6.2%)	24 (36.9%)	144 (27.1%)
	National Events	42 (13.9%)	20 (10.1%)	75 (92.6%)	35 (53.8%)	124 (23.3%)
	International Events	24 (7.9%)	12 (6.1%)	17 (21.0%)	34 (52.3%)	60 (11.3%)
Videography	User-Generated Style	261 (86.1%)	84 (42.4%)	236 (77.9%)	36 (55.4%)	350 (65.8%)
	Professional-User-Generated Style	30 (9.9%)	65 (32.8%)	10 (12.3%)	18 (27.7%)	112 (21.1%)
	Professional-Generated Style	8 (2.6%)	43 (21.7%)	8 (9.9%)	11 (16.9%)	70 (13.2%)
Visual Symbols	Internet-Meme Symbols	139 (45.9%)	17 (8.6%)	3 (3.7%)	6 (9.2%)	147 (27.6%)
	Traditional Symbols	110 (36.3%)	131 (66.2%)	9 (11.1%)	19 (11.1%)	201 (37.8%)
	International Symbols	17 (5.6%)	7 (3.5%)	13 (16.0%)	36 (55.4%)	78 (14.7%)
	‘Red ‘-Culture Symbols	23 (7.6%)	5 (2.5%)	65 (80.2%)	13 (20.0%)	53 (10.0%)
Music Genres	Popular Music	96 (31.7%)	38 (19.2%)	23 (28.4%)	19 (29.2%)	137 (25.8%)
	Traditional Music	52 (17.2%)	53 (26.8%)	4 (4.9%)	12 (18.5%)	93 (17.5%)
	Mix-Style Music	68 (22.4%)	74 (37.4%)	34 (42.0%)	14 (21.5%)	150 (28.2%)

Table 2. Distribution of content creation elements across ideological presentations

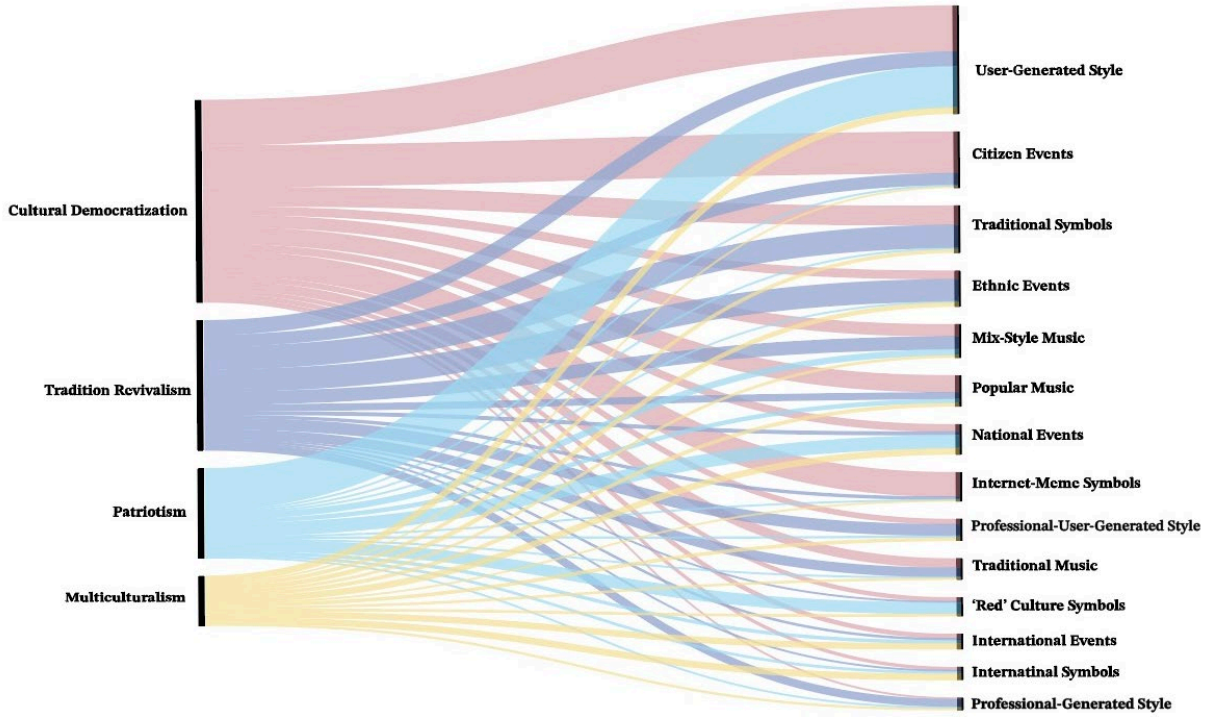


Figure 3. Sankey diagram of the content creation elements distribution

Discussion

Cultural democratization

Cultural democratization was the most frequent theme of ideological presentations (N=303; 56.96%), which refers to the advocacy of making cultural access, participation, and production available to a wider and more diverse population. Within the official SFVs published by ICHC, a non-formal, de-authoritative discourse environment was gradually constructed to democratize the ICH. The content allowed people to be active digital participants rather than passive information receivers of the culture. This looked for broader rights for communities and individuals by involving more stakeholders in heritage practices.

This ideological theme was most likely expressed through grassroots events. For instance, stories about the everyday lives of ordinary citizens, such as plots of marriage proposals or creating postcards for Teacher's Day, were frequently used by ICHC to engage a broad audience. In terms of videography, user-generated styles, including point-of-view and close-up shots, were employed to enhance visual appeal. The SFVs usually started with or were introduced by craftsmen or artisans, emphasizing the role of ICH artisans in heritage preservation and promotion practices. These kinds of interactive videography were hardly seen in previous ICH promotions by traditional media. Additionally, internet meme elements like popular animations or gifs, along with background music based on trending songs, were widely deployed in content creation. The incorporation of these elements made the SFV's content more layman-friendly and easier to understand, reducing the emphasis on authoritative official discourse in the promotion of ICH, while highlighting a more democratized engagement with popular culture.

Tradition revivalism

The second frequent ideological theme reflected in ICHC was tradition revivalism (N=198; 37.22%). In recent years, the importance of traditional culture for the improvement of China's soft power and the construction of the national ideological system has been increasingly emphasized. ICHC placed great emphasis on promoting traditional Chinese culture or philosophy. However, this

action did not refer to the blind praising of ancient values. We found that this ICHC leveraged SFVs to revitalize traditional customs or values of ICH, making them vibrant and relevant in today's political and societal landscape.

The SFVs categorizing this ideological theme often drew upon various ethnic events in their narratives, such as traditional festivals or folklore. This approach allowed the integration of diverse values from multiple ethnic traditions. For example, by reporting the dragon boat races during the *Duanwu* Festival, ICHC sought to convey current inheritances and interpretations of the ancient Chinese spirit of perseverance and determination. This also subtly implied the country's strength and prosperity. Regarding videographic presentation, SFVs featuring this theme adopted a more conservative style than others, frequently utilizing professional production formats like documentaries and official interviews. Traditional visual symbols, such as Spring Festival couplets and dragon stickers, were also extensively used in this content. However, we surprisingly observed that the background music in these SFVs showcased the fusion of traditional and popular elements. For instance, traditional instruments might be used to play popular music. The abovementioned SFV practices of ICH conveyed the ideological value of Chinese tradition revivalism, highlighting its important role in fostering ethnic and cultural identity.

Patriotism

Patriotism (N=81; 15.23%) was a key ideological theme embodied in the SFVs published by ICHC. These SFVs evoked a strong resonance of national identity among the audience, through presenting a sense of national pride or collectivist beliefs. Further, due to political and historical factors, the 'Red' culture was inextricably linked with patriotic beliefs in China (Huang et al., 2023), especially in the cyber context. The connection between ICH and the Communist Party of China (CPC) was constructed in these SFVs, to spread the ideology of patriotism.

In terms of narratives, these SFVs were often themed around national events, presenting the country's development milestones and heroic stories through ICH practice. For instance, an artisan used paper-cutting to commemorate important events like the national day. Another typical example was that during major disasters, such as the COVID-19 pandemic, much of the content connected to or conveyed the values of nationwide solidarity and collective resilience. In addition, the user-generated style of videography was primarily employed by ICHC, which showcased more spontaneous acts of patriotism by individuals, rather than state-driven initiatives. This significantly differed from the previous patriotism education or communication in mass media. Moreover, visual symbols associated with the identification of the country or CPC, such as the national flag, Tiananmen Square, or CPC emblem, were frequently incorporated to promote patriotic values. Notably, compared to other themes, SFVs featuring patriotic content often utilized mixed-style background music to evoke emotional responses. For instance, epic-style or 'Red culture' songs were performed using ICH instruments to heighten emotional impact.

Multiculturalism

Multiculturalism was also an important theme reflected in ICHC's ideological presentation (N=65; 12.22%). The values of respectfulness and inclusivity of diverse cultures were incorporated in many SFVs in ICHC. This constructed democratized discourse for the promotion of ICH from various ethnic groups. At the same time, ICHC also widely introduced the ICH from other countries and regions, such as African or Oceania tribes. This placed its official viewpoints of heritage in a broader context of globalization. Surprisingly, we also discovered that there was cross-cultural integration of ICH among ethnic groups, regions, and even countries in the ICHC's content. The folklore customs and artistic elements of various regions were dynamically represented, exchanged, and co-created, transcending the boundaries of individual groups.

Both national and international events were frequently selected as narrative topics. For example, during the 2022 Beijing Winter Olympics, ICHC published numerous works related to the event,

such as an ICH artisan using gourd carving to create the pattern of the panda mascot, *Bing Dwen Dwen*. In these types of SFVs, although user-generated style remained dominant, professional-generated, or professional-user-generated styles were more prominently featured compared to other SFVs with other ideological themes. For example, common filming techniques of documentaries and news reporting, such as wide-angle and high-angle shots, were frequently utilized, emphasizing a sense of objectivity by creating a distance from the subject (Berglund, 2007). In terms of visual elements, international symbols were extensively utilized in these SFVs. For instance, in response to the hosting of the 2023 Asian Games, the logo was transformed into a paper-cut design, becoming a significant visual motif in its presentation. Lastly, the content of this ideological theme did not focus on any genres of main or background music, which indirectly reflected their broader content positioning for a wider range of audience.

Conclusion

This study investigates the ideological presentations in the official promotion of ICH on SFV platforms. Through a multimodal content analysis of 532 SFVs published in ICHC, we first identify four primary ideological themes embodied in ICH promotion: (1) Cultural democratization, (2) Tradition revivalism, (3) Patriotism, and (4) Multiculturalism. Based on four proposed categories, namely *narratives*, *videography*, *visual symbols*, and *music genres*, we further investigate how these ideological themes were presented through various content creation elements of SFV.

Overall, the presentation of the ideological themes is not mutually exclusive. In addition, the application of various content creation elements of SFV is closely intertwined, collectively forming ideological presentations within the context of ICH promotion. In SFV platforms like Douyin, ICH promotion has increasingly reflected grassroots-centered narratives and user-generated videography. Elements of traditional and popular culture are integrated within this context. This fosters a relatively democratized discourse around ICH promotion, leading to a shift from conventional, authority-based perspectives towards a more inclusive and softened ideological expression.

Based on these findings, the contributions of this study can be articulated as follows. First, we provide new perspectives on what ideological themes are embodied in official SFVs promoting ICH and how they are presented through various content creation elements. Additionally, these findings offer insights into how current multimodal-oriented SFV platforms facilitate a decentralized and participative heritage practice (Giaccardi, 2012; Smith, 2017). This addresses the existing research gaps and contributes to both ideological and ICH research within SFV contexts. Moreover, drawing on Kress and van Leeuwen's (2020) social semiotic theory of multimodality, we developed a semiotic-based multimodal analytical template, which further extends the application of Serafini and Reid's (2023) multimodal content analysis framework within the SFV context. This template can be a valuable tool for future research in related areas. In terms of practical implications, our findings may benefit content creators, offering insights on how to effectively promote heritage and traditional culture to the public on digital media, particularly in SFV platforms.

A limitation of this study may point to directions for future research. We only collected data from one representative Douyin account for analysis. Although this allows for a richer and deeper investigation, it may also limit the generalizability of findings. Hence, future investigations could include more accounts for data collection. It is also considerable to conduct a comparative analysis between non-official and official content of ICH promotion, to gain complimentary findings.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable feedback.

About the authors

Ruoxi Liu is currently a Ph.D. candidate at the School of Art, Design and Media, Nanyang Technological University, Singapore. She holds an MA in material and visual culture from University College London, UK. Her research interests include social media, cultural heritage, digital humanities, and interactive design. Her email address is ruoxi001@e.ntu.edu.sg.

Haoran Qiu is the corresponding author of this paper. He is currently a Ph.D. candidate at Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. He holds an MSc. in knowledge management from Nanyang Technological University, Singapore. His research interests include social media, disinformation, information studies, and online behavior. His email address is haoran004@e.ntu.edu.sg.

References

Berglund, L. (2007). Chapter 9—Truth and Filmmaking—Objective and Subjective Camerawork. In R. Underwood (Ed.), Roll! (pp. 144–158). Focal Press. https://doi.org/10.1016/B978-0-240-80848-2.50044-8

Braun, V., & Clarke, V. (2012). Thematic analysis. In APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Cao, Z. (2022). Study of Creatively Popular Short Videos for Chinese Intangible Cultural Heritage. 103–106. https://doi.org/10.2991/assehr.k.220401.021

Chmil, H., Kuznietsova, I., Mishchenko, M., Oliynyk, O., & Demeshchenko, V. (2021). Intangible cultural heritage as a resource for consolidating modern Ukrainian society. Linguistics and Culture Review, 5(S4), 747–760. https://doi.org/10.21744/lingcure.v5nS4.1713

Cominelli, F., & Greffe, X. (2013). Chapter 20: Why and how intangible cultural heritage should be safeguarded. https://doi.org/10.4337/9780857931009.00033

Converse, P. E. (2006). The nature of belief systems in mass publics (1964). Critical Review, 18(1–3), 1–74. https://doi.org/10.1080/08913810608443650

Giaccardi, E. (2012). Introduction: Reframing heritage in a participatory culture. In Heritage and Social Media. Routledge. https://doi.org/10.4324/9780203112984

Huang, Z., Yang, Z., & Meng, T. (2023). National Identity of Locality: The State, Patriotism, and Nationalism in Cyber China. Journal of Chinese Political Science, 28(1), 51–83. https://doi.org/10.1007/s11366-022-09820-4

Kalay, Y., Kvan, T., & Affleck, J. (Eds.). (2007). New Heritage: New Media and Cultural Heritage. Routledge. https://doi.org/10.4324/9780203937884

Karampampas, P. (2021). (Re)inventing intangible cultural heritage through the market in Greece. International Journal of Heritage Studies, 27(6), 654–667. https://doi.org/10.1080/13527258.2020.1844277

Kress, G. (2009). Multimodality: A social semiotic approach to contemporary communication. routledge. https://doi.org/10.4324/9780203970034

Kress, G., & Leeuwen, T. van. (2020). Reading Images: The Grammar of Visual Design (3rd ed.). Routledge. https://doi.org/10.4324/9781003099857

Lähdesmäki, T. (2016). Politics of tangibility, intangibility, and place in the making of a European cultural heritage in EU heritage policy. International Journal of Heritage Studies, 22(10), 766–780. https://doi.org/10.1080/13527258.2016.1212386

Lenzerini, F. (2011). Intangible Cultural Heritage: The Living Culture of Peoples. European Journal of International Law, 22(1), 101–120. https://doi.org/10.1093/ejil/chr006

Lichtheim, G. (1965). The Concept of Ideology. History and Theory, 4(2), 164–195. https://doi.org/10.2307/2504150

Macdonald, S. (2003). 20th-Century Heritage: Recognition, Protection and Practical Challenges. Heritage at Risk, 223–229. https://doi.org/10.11588/hr.2003.0.21220

Massing, K. (2018). Safeguarding intangible cultural heritage in an ethnic theme park setting – the case of Binglanggu in Hainan Province, China. International Journal of Heritage Studies, 24(1), 66–82. https://doi.org/10.1080/13527258.2017.1362571

Petronela, T. (2016). The Importance of the Intangible Cultural Heritage in the Economy. Procedia Economics and Finance, 39, 731–736. https://doi.org/10.1016/S2212-5671(16)30271-4

Qiu, Q., Zuo, Y., & Zhang, M. (2022). Intangible cultural heritage in tourism: Research review and investigation of future agenda. Land, 11(1), 139. https://doi.org/10.3390/land11010139

Rodrigues, T. A. (2023). Cultural Experiences of Brazil in Africa: The Management of Intangible Cultural Heritage in the Context of Soft Power. In R. Christofoletti (Ed.), Soft Power and Heritage (pp. 221–233). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-41207-3_13

Ryan, K. M. (2018). Vertical video: Rupturing the aesthetic paradigm. Visual Communication, 17(2), 245–261. https://doi.org/10.1177/1470357217736660

Schreiber, H. (2017). Intangible cultural heritage and soft power–exploring the relationship. International Journal of Intangible Heritage, 12, 44–57. http://doi.org/10.35638/ijih.2017..12.005

Schreier, M. (2012). Qualitative Content Analysis in Practice. SAGE Publications, Inc. https://doi.org/10.4135/9781529682571

Serafini, F., & Reid, S. F. (2023). Multimodal content analysis: Expanding analytical approaches to content analysis. Visual Communication, 22(4), 623–649. https://doi.org/10.1177/1470357219864133

Smith, L. (2006). Uses of Heritage. Routledge. https://doi.org/10.4324/9780203602263

Smith, L. (2017). Heritage, Identity and Power. In H.-H. M. Hsiao, Y.-F. Hui, & P. Peycam (Eds.), Citizens, Civil Society and Heritage-making in Asia (pp. 15–39). ISEAS Publishing. https://doi.org/10.1355/9789814786294-003

Sullivan, A. M. (2015). Cultural Heritage & New Media: A Future for the Past. John Marshall Review of Intellectual Property Law, 15(3), 604–646. https://repository.law.uic.edu/ripl/vol15/iss3/11/

Tavares, D. S., Alves, F. B., & Vásquez, I. B. (2021). The relationship between intangible cultural heritage and urban resilience: A systematic literature review. Sustainability, 13(22), 12921. https://doi.org/10.3390/su132212921

Taylor, J., & Gibson, L. K. (2017). Digitisation, digital interaction, and social media: Embedded barriers to democratic heritage. International Journal of Heritage Studies, 23(5), 408–420. https://doi.org/10.1080/13527258.2016.1171245

Taylor, M. N. (2009). Intangible heritage governance, cultural diversity, ethno-nationalism. Focaal, 2009(55), 41–58. https://doi.org/10.3167/fcl.2009.550104

UNESCO. (2022). Basic Texts of the 2003 Convention for the Safeguarding of the Intangible Cultural Heritage. https://ich.unesco.org/en/convention

UNESCO. (2024). UNESCO - Browse the Lists of Intangible Cultural Heritage and the Register of good safeguarding practices. https://ich.unesco.org/en/lists

Van Daele, T., Iyer, A., Zhang, Y., Derry, J. C., Huh, M., & Pavel, A. (2024). Making Short-Form Videos Accessible with Hierarchical Video Summaries. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–17. https://doi.org/10.1145/3613904.3642839

Vecco, M. (2010). A definition of cultural heritage: From the tangible to the intangible. Journal of Cultural Heritage, 11(3), 321–324. https://doi.org/10.1016/j.culher.2010.01.006

Vrdoljak, A. F. (2005). Minorities, cultural rights, and the protection of intangible cultural heritage. European Society for International Law. Research Forum on International Law Contemporary Issues. Geneva: Graduate Institute of International Studies. http://www.esil-sedi.org/english/pdf/Vrdoljak09-05.pdf

Wang, H., Zhao, M., Hu, W., Ma, Y., & Lu, Z. (2024). Critical Heritage Studies as a Lens to Understand Short Video Sharing of Intangible Cultural Heritage on Douyin. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–21. https://doi.org/10.1145/3613904.3642138

Wang, Y. (2020). Humor and camera view on mobile short-form video apps influence user experience and technology-adoption intent, an example of TikTok (DouYin). Computers in Human Behavior, 110, 106373. https://doi.org/10.1016/j.chb.2020.106373

Wang, Y. (2021). Content characteristics and limitations of original short video based on depth data. Journal of Physics: Conference Series, 1881(4), 042070. https://doi.org/10.1088/1742-6596/1881/4/042070

Xu, Y., Tao, Y., & Smith, B. (2022). China's emerging legislative and policy framework for safeguarding intangible cultural heritage. International Journal of Cultural Policy, 28(5), 566–580. https://doi.org/10.1080/10286632.2021.1993838

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Professional identity scale for information professionals: an instrument development

Fatih Oguz
DOI: https://doi.org/10.47989/ir30iConf47314

Abstract

Introduction. There is no validated and reliable instrument available for information professionals to measure their professional identity development. As the profession goes through a transformation and recalibrates itself to stay relevant, enhance training opportunities and programs to promote a strong professional identity, and better address the needs and expectations of users, having a validated and reliable instrument to assess the professional identity of the professionals in the field will help address this gap.

Method. A quantitative approach was used to collect data about LIS professional identity to develop a reliable scale. A non-probability sampling strategy was employed to recruit early career LIS professionals.

Analysis. Data were analysed using exploratory factor analysis approach.

Results. The proposed instrument has high internal reliability and captures aspects of the LIS professional identity. Results of the exploratory factor analysis indicated that five areas or subscales can assess the LIS professional identity.

Conclusion. The results of the study supported that proposed constructs were mostly consistent with LIS professional identity scale factors. Though there is extensive research in professional identity development of LIS professionals, this is a first step to developing a scale to measure LIS professional identity.

Introduction

About 400 publications have been published in the scholarly and professional literature with the words '*future of libraries*' or '*future of librarianship*' in any of their text fields since 2016. This number (~3500) was significantly higher when these search terms were used in Google Scholar for the same period as it has a broad coverage of databases and repositories. The same search yielded more than 500 publications in 2016 for the previous 5-year period (Dorner et al., 2017). Such an interest in the future of the field in the community is no coincidence as the library and information science (LIS) profession has been going through a major transformation period (Fraser-Arnott, 2019) due to emergence of new digital technologies that dramatically improve access to information for the average user (Fraser-Arnott, 2019; IFLA Trend Report, 2019; Witt & Smith, 2019) and, as a result, transformations in the global information economy (IFLA Trend Report, 2019).

Professional identity is generally defined as one's interpretation of self as a professional in connection with the personal and professional experiences, attitudes, and aspirations (e.g., Cheung, 2008; Whyte, 2002). Definitions of professional identity in the literature appear to share three common elements: (1) self-labelling as a professional, (2) integration of skills and attitudes as a professional, and (3) a perception of context in a professional community (Fraser-Arnott, 2019; Gibson et al., 2010). Therefore, one's professional identity needs to be regarded as a fluid construct as it can change over time through learning, experiences, and interactions with others within the profession and beyond (Cheung, 2008; Dorner et al., 2017; Fraser-Arnott, 2019; Pierson et al., 2019).

The LIS profession goes through transformations as a response to challenges and opportunities from the external environment (e.g., digital evolution) and within (e.g., workforce demographics) which in turn requires changes to the professional identities of the individuals working in the field (Fraser-Arnott, 2019). There is no validated and reliable instrument available in LIS to measure professional identity development of information professionals. As the profession goes through a transformation and recalibrates itself to stay relevant, enhance training opportunities and programs to promote a strong professional identity, and better address the needs and expectations of users, having a validated and reliable instrument to assess the professional identity of the professionals in the field will help address this gap.

Literature review

Socialization within a professional community plays a key role in construction of professional identities (Croxton, 2015; Dobrow & Higgins, 2005; Dorner et al., 2017; Fraser-Arnott, 2019), the theories of social identity (Tajfel, 1978) and communities of practice (Wenger, McDermott, et al., 2002) are selected as a theoretical framework to inform the development of a professional identity development scale for information professionals.

According to the social identity theory, individuals drive one part of their total identity, their social identity, from the group or groups that they belong to (Tajfel, 1978). Social identity is '*that part of an individual's self-concept which derives from knowledge of membership in a social group (or groups) together with the value or emotional significance attached to that membership*' (Tajfel, 1978, p. 63). The social group's features, norms, values, and expectations determines not only the strength and content of social identity but also emotions and behaviours. A group of people who engage in joint learning activities, build relationships, and help each other regularly in pursuing their interests in the domain is described as a community (Wenger, Mcdermott, et al., 2002) and continuity in their interactions allows them to develop a sense of belonging, identity, and commitment.

Professional identity cannot be treated as a fixed attribute (Cheung, 2008) as it is informed and shaped by a variety of factors that influence how one perceives themselves as a professional. Formal education and training is seen as a foundational step to building a sense of professionalism (Hussey & Campbell-Meier, 2016; Sare et al., 2012; Tan et al., 2017; Trede et al., 2012) as students

are introduced to knowledge, skills, ethical expectations, and professional values of the field. Through knowledge of the profession (Tan et al., 2017) and practical experiences (Tan et al., 2017; Trede et al., 2012) in the field such as internships or prior work experience (Bird et al., 2011; Croxton, 2015), students align themselves with the expectations of the profession as such experiences also allow students expand their social networks in their field (Tomlinson & Jackson, 2021). Professional identity is also cultivated through interactions with peers, mentors, and professional networks (Croxton, 2015; Hussey & Campbell-Meier, 2016; Pierson et al., 2019). Engagement with professional associations (Pierson et al., 2020; Woo & Henfield, 2015) promote their socialization into the field through opportunities such as networking (Dobrow & Higgins, 2005), professional development, and advocacy. LIS profession is grounded in a set of core values and ethical principles including intellectual freedom, public good, and equitable access (e.g., *Core Values of Librarianship* - ALA, 2024; *Our Vision and Mission* - IFLA, n.d.). Internalization of such values and ethical principles not only provide a foundation for professionals perform their roles and responsibilities in the field (Woo & Henfield, 2015) but also allow them to align themselves with their peers, enhance their sense of belonging to the community (Wenger, Mcdermott, et al., 2002), and strengthen their professional identity (Adams et al., 2006; Pierson et al., 2020).

Method

The study employed a quantitative approach to develop an instrument to measure professional identity development of information professionals. A framework proposed by Pierson et al. (2019) included 14 characteristics relevant to professional identity development of information professionals informed the instrument development. A mapping of values, ethics, and competences statements of major professional associations including the American Library Association, the American Alliance of Museums, and the International Federation of Library Associations and Institutions developed to inform the instrument development. Although professional identity development scales developed in other helping professions including the professional identity scale for in-service teachers (Cheung, 2008) and the professional identity scale in counselling (Woo et al., 2018; Woo & Henfield, 2015) are discipline specific, they served as a model and appropriate measures are adapted. Professional identity in LIS domain is conceptualized as the product of six areas of an LIS professional's ability to

(1) demonstrate knowledge of the profession (*Knowledge of the Profession*) [7 items],
(2) characterize the values that underlines the profession (*Philosophy of the Profession*) [7 items],
(3) show an understanding of expertise required and professional roles (*Professional Roles and Expertise*) [9 items],
(4) endorse attitudes toward the profession and oneself (*Attitude*) [9 items],
(5) socialize with other professionals in a range of capacities (*Engagement Behaviours*) [11 items], and
(6) manage outward perception of the profession (*Perceptions*) [5 items].

A total 63 item questions were initially developed for these constructs.

Expert review

Eleven expert reviewers were invited to perform content validation of the instrument and provide feedback. These experts included practitioners as well as LIS faculty members with expertise on the subject. Of the eleven experts, eight reviewed the instrument and provided feedback. Reviewers are provided with definitions, a list of literature references that informed the development of each construct, and a set of questions proposed to measure each construct. Reviewers were asked to rate each question with a seven-point Likert scale ranging from *Strongly agree* to *Strongly disagree*. A comment area was included for each construct and another comment area for the instrument for reviewers to share their feedback. The questionnaire was disseminated

to expert reviewers online. Reviewers were generally in agreement about the proposed constructs. As per reviewer feedback, several questions were revised or removed. A total of 47 questions for six constructs were included in the final instrument.

Instrument and data analysis

Nine demographics questions such as graduation year from a Master of LIS (MLIS) program, employment status, work experience, and race/ethnicity, and gender were included in addition to 47 questions about six constructs. There was no suitable sampling frame for LIS professionals available. Therefore, the instrument was disseminated online via various alumni mailing lists, Reddit, Facebook groups, and direct emailing. Career transitions require changes to one's professional identity as new roles demand new skills, perspectives, and behaviours (Dobrow & Higgins, 2005; Ibarra, 1999). Therefore, transition of students from school to work presents a unique opportunity to study the professionals' identity development (Dobrow & Higgins, 2005). Recent graduates of MLIS programs (Fall 2018 or later) in the US were invited to participate in the study. Of 403 responses were received, 311 responses were used for data analysis.

Participants were predominantly identified as White (%85.9, n=267) and female (64.6%, n=201). Almost all the respondents were employed either full-time (82%, n=255), part-time (13.5%, n=42), or self-employed (4.2%, n=13) at the of data collection. About half of respondents (52.4%, n=163) had a work experience of 5 years or less in LIS domain.

		Factor				
	Item	1	2	3	4	5
	Factor 1: Professional Integrity					
1	Upholding the principle of intellectual freedom and users' right to privacy in accessing information is essential.	0.765				
2	Regardless of different roles an LIS professional may hold, key goals include facilitating access to information and serving our users.	0.714				
3	My profession is passionate about connecting users with information that supports a wide range of needs, including education, research, intellectual inquiry, and recreational/entertainment interests.	0.705				
4	A commitment to provide the highest service levels to promote informed, connected, educated, and empowered communities is important to contribute to the public good.	0.704				
5	Equitable access to information regardless of technology, format, or delivery methods is one of the core values in my profession.	0.667				
6	I value the advancement and the future of my profession.	0.662				
7	I value various professional roles (e.g., educator, librarian, and information professional) that an LIS professional can hold.	0.658				
8	My profession provides unique and valuable services to society.	0.634				
9	Innovation and adaptability to meet evolving user needs and technological advancements in the field is integral to the LIS philosophy.	0.633				
10	I am knowledgeable of ethical responsibilities and professional standards relevant to my roles.	0.628				
11	I believe an LIS professional should value the importance of advocacy for the profession.	0.626				
12	Advocacy for libraries, information sources and services, and library users is emphasized in the LIS philosophy.	0.620				
13	I believe an LIS professional should value the importance of advocacy for the community that the person serves.	0.532				
	Factor 2: Engagement Behaviours					
14	I follow up with theoretical, practical, and technical advancement in my profession by keeping up with literature (e.g., professional journals, books) in the field.		0.704			
15	I have contributed to research in LIS (e.g., by being interviewed, taking surveys, or conducting research).		0.689			
16	I engage in or seek voluntary service opportunities in my professional community (e.g., professional associations, interest networks, committees, or community service).		0.671			
17	I regularly communicate with a mentor/mentee who is interested in my/their professional development.		0.654			
18	I am an active member of my professional community (e.g., participating in conferences, workshops, or webinars) every year.		0.649			

19	I have presented/published to contribute to the field's knowledge base or share insights/experiences.	0.623				
20	I seek feedback/consultation from professional peers/colleagues as a form of professional development.	0.601				
21	I have membership(s) in LIS professional associations (e.g., local, regional, state, national, or international).	0.593				
22	I advocate for my profession by participating in activities associated with legislation, law, or policy on behalf of the profession.	0.569				
23	I participate in ongoing discussions with colleagues about identity and the vision of the profession.	0.562				
	Factor 3: Knowledge of the Profession					
24	I am familiar with accreditation organizations (e.g., American Library Association's [ALA] Committee on Education for accrediting educational programs in North America) and their standards for professional preparation.		0.685			
25	I am knowledgeable of professional LIS publications (e.g., American Libraries) publication(s) relevant to my specialty area and their purposes in the profession.		0.674			
26	I am knowledgeable about ethical guidelines (e.g., codes of ethics/standards of practice) in LIS.		0.656			
27	I am familiar with laws and regulations (e.g., copyright, licensure) related to my profession.		0.653			
28	I am familiar with professional LIS associations (e.g., international, national, regional, or local associations) and their roles in the profession.		0.616			
29	I am informed about the important events and milestones (e.g., establishment of public libraries, introduction of standards and cataloguing systems, role of academic libraries, rise of digital libraries, and open access) in Library and Information Science (LIS) history.		0.605			
30	I am familiar with the requirements to work as a professional (e.g., ALA-accredited MLIS degree, state-specific requirements) in my field.		0.593			
31	I completed my professional training and standard education to perform my duties in my roles.		0.517			
	Factor 4: Attitude					
32	I am satisfied with my work and professional roles.			0.775		
33	I recommend my profession to those who are searching for a new career related to helping professions.			0.712		
34	I share my positive attitude towards the profession when working with people in other fields.			0.665		
35	My profession fosters a mindset of collaboration, empathy, cultural humility, ethical conduct, and continuous learning.			0.637		
36	I am confident that there will be positive outcomes of my work and services.	0.451		0.535		
	Factor 5: Perceptions					
37	I am confident in my ability to address stereotypes/misconceptions about my profession.					0.682

38	I regularly communicate with others about the roles of LIS professionals in society.					0.661
39	I would like to more involved in developing the public perception of the profession.					0.654
	Eigenvalues	9.586	5.990	2.185	1.471	1.363
	% of variance	17.348	12.083	10.660	7.630	5.088
	Cumulative %	17.348	29.431	40.091	47.720	52.808

Table 1. Factor loadings and communalities of exploratory factor analysis

The Cronbach's alpha statistic was estimated as 0.90 indicating a high level of internal consistency. In this analysis, the KMO measure was 0.90, suggesting enough of a satisfactory factor analysis to proceed, and the Bartlett's test of sphericity being statistically significant ($p<0.001$). The results of an exploratory factor analysis on LIS professional identity scale items identified five factors which explained 52.8% of the total variance as shown in Table 1. The criteria recommended by Woo and Henfield (2015) were adopted to determine an appropriate factor structure for the instrument. Thirty-nine items consisting of five factors remained from the original instrument after eliminating eight items that did not meet the criteria.

Discussions

The results suggest that the proposed instrument has high internal reliability and captures aspects of the LIS professional identity. Though professional identity in LIS domain was originally conceptualized as the product of six areas of an LIS professional's ability to perform, results of the exploratory factor analysis indicated that five areas or subscales can assess the LIS professional identity as shown in Table 1. The most central aspect in the scale was the *Professional Integrity* factor. This subscale comprised of 13 items from three areas: philosophy of the profession, professional roles and expertise, and attitude. Although these items were originally conceptualized to be part of three separate areas, they were grouped into a single factor. Since these items are related to professional values, ethics, and advocacy which are integral to how LIS professionals approach their responsibilities, this subscale was named as *Professional Integrity*. Internalization of professional values, ethics, and advocacy ensures that professionals consistently uphold the principles and responsibilities of their profession, informing their actions and decisions in a way that benefits society and advances the profession. This alignment cultivates a sense of purpose, belonging (Wenger, Mcdermott, et al., 2002), and commitment to their professional role (Adams et al., 2006; Pierson et al., 2020).

The second most important subscale emerged as the *Engagement Behaviours*. This area measures an LIS professional's capability to perform professional actions and activities, and socialization into the field. As expected, proposed items for this area were grouped into a single factor. Participation in professional associations and engagement with the community play a critical role in strengthening one's professional identity through interaction with their peers and community (Croxton, 2015; Dobrow & Higgins, 2005; Woo & Henfield, 2015).

The third factor emerged as *Knowledge of the Profession*. Items that were originally conceptualized in this area emerged under this factor with the addition of an item from the original professional roles and expertise area. Since the additional item was about professional training and preparation which aligned well with the area's focus on measuring an LIS professional's basic knowledge and understanding of the profession, the name of the area was retained. Respondents agreed that education and training play a critical role in building and strengthening one's professional identity (Hussey & Campbell-Meier, 2016; Sare et al., 2012; Tan et al., 2017; Trede et al., 2012).

The fourth factor emerged from the analysis was the *Attitude*. Proposed items in this area were grouped into a single factor with an additional item from professional roles and expertise. The

added item was the professional's self confidence in their work and services. Since the original construct measures an LIS professional's perspective toward the profession and oneself and the added item was related to one's attitude towards their own work, the name of the subscale was retained. One's perspective towards the profession is influenced by a combination of factors including personal and social influences (Fraser-Arnott, 2019) including education and training, professional community, work experiences, and perception of the profession by others.

The last factor emerged from the data was the *Perception*. This area focused on capturing an LIS professional's perception of the profession (stereotypes/misconceptions) by others. Three items that were originally conceptualized in this area were grouped into a single factor. Perception of the profession and professionals by others is an important factor in informing LIS professional identity as professional often ground their identity in services that they provide (Fagan et al., 2021; Hicks, 2016).

Limitations

The use of a non-probability sampling strategy to recruit participants limits generalizability of the results to a larger population. A confirmatory factor analysis needed to increase the generalizability of the results. Additionally, the instrument may have a North American bias in questions related to accreditation as such an external accreditation structure may apply to other parts of the world.

Conclusion

The results of the study supported that proposed constructs were mostly consistent with LIS professional identity scale factors. Though there is extensive research in professional identity development of LIS professionals, this is a first step to developing a scale to measure LIS professional identity. Such a scale can be useful tool for LIS education and training programs to assess the impact of their activities on participants' professional identity development. LIS students and early career professionals may also use such a scale to assess their own preparation for the profession by identifying their strengths and weaknesses.

About the author

Fatih Oguz is Associate Professor in Department of Library, Information, and Research Sciences at the University of the North Carolina at Greensboro, USA. He received his Ph.D. from University of North Texas, and his research focuses on the interactions between people and technology, with an emphasis on the impact of information technologies on human communication behaviour and social life. He can be contacted at f_oguz@uncg.edu.

References

Adams, K., Hean, S., Sturgis, P., & Clark, J. M. (2006). Investigating the factors influencing professional identity of first-year health and social care students. *Learning in Health and Social Care*, 5(2), 55–68. https://doi.org/10.1111/j.1473-6861.2006.00119.x

Bird, N. J., Chu, C. M., & Oguz, F. (2011, August 13). *Four 'I's of Internships for the New Information Society: Intentional, Interconnected, Interdisciplinary and International*. 77th IFLA General Conference and Assembly, San Juan, Puerto Rico. http://libres.uncg.edu/ir/uncg/listing.aspx?styp=ti&id=8373

Cheung, H. Y. (2008). Measuring the professional identity of Hong Kong in-service teachers. *Journal of In-Service Education*, 34(3), 375–390. https://doi.org/10.1080/13674580802003060

Core Values of Librarianship—ALA. (2024, January). [Text]. Advocacy, Legislation & Issues. https://www.ala.org/advocacy/advocacy/intfreedom/corevalues

Croxton, R. A. (2015). Professional Identity Development Among Graduate Library and Information Studies Online Learners: A Mixed Methods Study. *Community & Junior College Libraries*, 21(3/4), 125–141. https://doi.org/10.1080/02763915.2016.1190998

Dobrow, S. R., & Higgins, M. C. (2005). Developmental networks and professional identity: A longitudinal study. *Career Development International*, 10(6/7), 567–583. https://doi.org/10.1108/13620430510620629

Dorner, D., Campbell-Meier, J., & Seto, I. (2017). Making sense of the future of libraries. *IFLA Journal*, 43(4), 321–334. https://doi.org/10.1177/0340035217727554

Fagan, J. C., Ostermiller, H., Price, E., & Sapp, L. (2021). Librarian, Faculty, and Student Perceptions of Academic Librarians: Study Introduction and Literature Review. *New Review of Academic Librarianship*, 27(1), 38–75. https://doi.org/10.1080/13614533.2019.1691026

Fraser-Arnott, M. A. (2019). Evolving practices and professional identity: How the new ways we work can reshape us as professionals and a profession. *IFLA Journal*, 45(2), 114–126. https://doi.org/10.1177/0340035218810960

Gibson, D. M., Dollarhide, C. T., & Moss, J. M. (2010). Professional Identity Development: A Grounded Theory of Transformational Tasks of New Counselors. *Counselor Education and Supervision*, 50(1), 21–38. https://doi.org/10.1002/j.1556-6978.2010.tb00106.x

Hicks, D. (2016). Advocating for Librarianship: The Discourses of Advocacy and Service in the Professional Identities of Librarians. *Library Trends*, 64(3), 615–640.

Hussey, L. K., & Campbell-Meier, J. (2016). Developing professional identity in LIS? *Education for Information*, 32(4), 343–357. https://doi.org/10.3233/EFI-160981

IFLA Trend Report: Riding the Waves or Caught in the Tide? (2019). https://trends.ifla.org/insights-document

Our Vision and Mission—IFLA. (n.d.). IFLA. Retrieved April 2, 2024, from https://www.ifla.org/vision-mission/

Pierson, C. M., Goulding, A., & Campbell-Meier, J. (2019). An integrated understanding of librarian professional identity. *Global Knowledge, Memory and Communication; Bingley*, 68(4/5), 413–430. http://dx.doi.org/10.1108/GKMC-01-2019-0008

Pierson, C. M., Goulding, A., & Campbell-Meier, J. (2020). Professional Identity as Gateway to Critical Practices: Identity Negotiations of Public Librarians in New Zealand with Implications for LIS Education and Practice. *Journal of the Australian Library and Information Association*, 69(3), 286–301. https://doi.org/10.1080/24750158.2020.1777636

Sare, L., Bales, S., & Neville, B. (2012). New Academic Librarians and Their Perceptions of the Profession. *Portal: Libraries and the Academy*, 12(2), 179–203. http://dx.doi.org/10.1353/pla.2012.0017

Tajfel, H. (1978). Social categorization, social identity and social comparisons. In H. Tajfel (Ed.), *Differentiation between social groups* (pp. 61–76). Academic Press.

Tan, C. P., Van der Molen, H. T., & Schmidt, H. G. (2017). A measure of professional identity development for professional education. *Studies in Higher Education*, 42(8), 1504–1519. https://doi.org/10.1080/03075079.2015.1111322

Tomlinson, M., & Jackson, D. (2021). Professional identity formation in contemporary higher education students. *Studies in Higher Education*, 46(4), 885–900. https://doi.org/10.1080/03075079.2019.1659763

Trede, F., Macklin, R., & Bridges, D. (2012). Professional identity development: A review of the higher education literature. *Studies in Higher Education*, 37(3), 365–384. https://doi.org/10.1080/03075079.2010.521237

Wenger, E., McDermott, R., & Snyder, W. M. (2002). A *Guide to Managing Knowledge: Cultivating Communities of Practice*. Harvard Business School Press.

Wenger, E., Mcdermott, R., & Snyder, W. M. (2002). A *guide to managing knowledge: Cultivating communities of practice*. Harvard Business School Press.

Whyte, W. H. (2002). *The organization man*. University of Pennsylvania Press.

Witt, S., & Smith, K. (2019). Libraries in times of crisis. *IFLA Journal*, 45(1), 3–4. https://doi.org/10.1177/0340035219830549

Woo, H., & Henfield, M. S. (2015). Professional Identity Scale in Counseling (PISC): Instrument Development and Validation. *Journal of Counselor Leadership and Advocacy*, 2(2), 93–112. https://doi.org/10.1080/2326716X.2015.1040557

Woo, H., Lu, J., & Bang, N. (2018). Professional Identity Scale in Counseling (PISC): Revision of Factor Structure and Psychometrics. *Journal of Counselor Leadership and Advocacy*, 5(2), 137–152. https://doi.org/10.1080/2326716X.2018.1452078

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Stewardship of digital language archives: training development and testing through collaboration of information scientists, linguists, and communities

Oksana L. Zavalina, Alexandra O'Neil, and Shobhana L. Chelliah

DOI: https://doi.org/10.47989/ir30iConf47311

Abstract

Introduction. Specialized repositories aggregating digital data that focus on languages (of indigenous groups, refugees, etc.) are known as digital language archives (DLA). DLA rapid growth is galvanized by language documentation efforts supported by funding agencies. Recent studies examine DLA user needs and explore support of these needs in digital repositories. Training of LAM professionals in curating and managing DLA to support user needs is emerging, with no research yet into its effectiveness.

Method. Our federally funded interdisciplinary project creates and tests the 1st US LAM graduate course with DLA focus. The curriculum development builds on team's experience with DLA, available DLA user studies, and collaborations of LAM and linguistics researchers with language communities. We report on the current state of curriculum development, discuss analysis results and future steps.

Analysis. Quantitative and qualitative analyses of data collected through course-level and module-level surveys, and content analysis.

Results. Overall, the project-developed training materials are effective in developing learning objectives. Areas for improvement are identified.

Conclusion. The team is refining the LAM curriculum and developing language community DLA training workshops based on student evaluation results, collecting community feedback for future analysis. The project is expected to positively impact LAM education and DLA users' experiences.

Introduction

A digital language archive (DLA) is an archive that includes born-digital and/or digitized language resources in any language(s) or dialect(s), a searchable repository in which resource discovery is powered by metadata. The collections comprising DLAs are often created with the goal of preserving, or documenting, a language, especially an endangered language. Once documented, these materials assist the revitalization efforts of the language(s) in question (Hinton, 2001). Thus, items selected for these collections demonstrate language use in various settings and through different media and file formats (Chelliah, 2021). Some of the common item types in DLAs are also collected by broader-scope digital repositories stewarded by cultural heritage institutions, and include audio-visual recordings of oral histories, performances, interviews, text resources: books, articles, letters, diaries, etc. However, DLA also include a significant proportion of resource types that are not found in other digital repositories: wordlists, field notes, AI-generated and Large Language Models related content, corpus publications essential for computational linguistics research; linguistic annotations, transcriptions, and software (Bird & Simons, 2003).

Literature review

In the past, language communities, despite often being the source of language archive materials, were typically not involved in the creation or maintenance of archives (Henke & Berez-Kroeker, 2016). This practice privileged linguistics and anthropology researchers, missionaries, and colonial administrators as creators of archival records (i.e., items and collections), ascribing provenance to them over the individuals and communities depicted in the archival records (Jones, 2019). However, like other archives, DLAs undergo significant changes and evolve to meet the requirements of the archival paradigm shift toward community-oriented ways of archival thinking and participatory methodologies (Cook, 2013; Genovese, 2016; Rolan, 2017). For DLAs, that means that source language communities of the archival records take an increasingly active part in decision-making regarding donation, appraisal, preservation, metadata, access, dissemination, etc. (Burke, 2021; Dale et al., 2022; Harris et al., 2019; Olson, 2023; Pugh & Christen, 2011; Roeschley & Kim, 2019).

Information science and linguistics researchers explore the archive user needs of language community representatives and scholars. Developing understanding of these needs is important for improving access as they guide user interactions with and expectations towards DLA. Some of the user needs tend to focus on the services offered by the DLA and include stream-able audio and video, user interface accessibility on mobile devices, as well as machine-readable text files and bulk download availability necessary for computational linguistics research (e.g., Burke et al., 2022). Endangered language community members also report needs – related to supporting the revitalization of their languages – that inform collection development decisions as language communities representatives call for inclusion of specific kinds of resources beyond the obvious choice of dictionaries, audio and video files: textbooks and teaching aids covering different disciplines, folktales and storybooks for children, resources on religion and culture, and overall attractive items that capture attention (Burke, 2023). Research demonstrates that information needs of DLA users are often not met, and that one of the key reasons for this is the lack of academic instruction for information professionals on supporting the stewardship and digital curation of these collections (Al Smadi et al., 2016; Wasson, Holton, & Ross, 2016; Zavalina & Chelliah, 2021).

While there is an overlap in the principles and practices of broader digital archiving, digital archiving of linguistics research data, and community digital archiving focusing on language, there are substantial differences in the goals, scale, content, challenges, ethical considerations, and impact. One example is that types of items unique to language archives and specific attributes of these resources, as well as complex relations between these items must be represented differently in bibliographic metadata, archive's navigation, and controlled vocabularies to meet DLA user information needs (e.g., Aljalahmah & Zavalina, 2023; Paterson, 2024; Zavalin, 2023). Greater focus

is needed on authorship and intellectual rights (e.g., with AI-generated data), as well as on ownership, archiving spaces, and mutually beneficial relations between communities and partnering researchers and institutions, identifying, and disseminating best practices, and developing collaborations to support community archiving and to engage communities in stewardship of their DLA collections housed by LAMs (Zavalina & Chelliah, 2021; Chelliah, 2023).

University of North Texas (UNT) educators experimented with adding DLA training for LAM students as an individual DLA-focused module in the advanced metadata course. Preliminary results of DLA-focused learning module effectiveness indicated that during 2021-2023 testing, 75% to 90% students attained the target cumulative score of 85% or more on module assignments and activities (Zavalina, 2023). Student survey data on teaching effectiveness however were collected and analysed only at the course-level, not for the DLA-focused module. Realization of the need for a more robust in-depth and open-access DLA training for LAM students led to successful proposal to the US Institute or Museum and Library Services (IMLS) for addressing this need.

Expanding LAM workforce training

To meet this demand, our project's interdisciplinary team develops curriculum with a strong experiential component to educate information professionals in the archiving and curation of resources that provide the means to revitalize community memory and language. The project is producing learning materials intended to be useable in courses offered in LIS and Archival university programs, LAM on-the-job trainings, and archiving workshops for members of indigenous, immigrant & refugee communities who are documenting their community heritage.

Project-developed course, data collection and analysis

Our team developed the online graduate course that consists of 4 modules and introduces LAM students to how the language archives have evolved and function today, to needs of the end-users and depositors of these archives, to material types that DLAs collect, specifics of describing the unique attributes of these complex resources in metadata, and digital curation of these materials. This course also introduces students to working with language communities on collection stewardship, ethical and effective dissemination of language archive resources, and evaluation of DLAs. The course content, developed since 2023, successfully went through university-administered assessment of online course pedagogy, copyright/trademark, and ADA accessibility compliance. It was first offered as a section of an existing information resources and services for special clienteles seminar to UNT LAM graduate students during June 24-July 26, 2024, with 11 IMLS-funded tuition stipends, and 23 students completing the course.

At multiple points of time, we collected and analysed various data on effectiveness of the course and its individual learning modules. Student success in completing the tasks of the 4 module assignments was used in the course evaluation: we analysed range, average, standard deviation of the scores attained by student submissions and compared these across modules. At the end of semester, the team obtained anonymous student survey data to evaluate effectiveness in meeting each of the 8 broader course-level learning objectives. Data collection also included anonymous pre-test and post-test surveys to evaluate each of the 4 learning modules, with questions designed to capture indications of perceived progress made by the student and based on the much more specific module-level learning objectives (a total of 23, ranging 4-7 per module). Finally, our team analysed quantitatively and qualitatively the results of the university-administered online teaching effectiveness survey, after these results were released to the instructor in August 2024.

Course evaluation findings and discussion

Results of assignment submissions evaluation demonstrate that 95.65% of students consistently produced high-quality work, earning 80% or more of possible points. All (100%) students attained the cumulative average score (for all 4 assignments) of 85% or more, higher than in Zavalina (2023). The average score attained ranged from assignment to assignment between 18.26-19.75 on the 20-

point scale (91%-99%), with the highest observed for the module 2 comparative discussion assignment and the lowest for the module 4 practical exercise. The highest standard deviation in the student scores (2.464) was observed in module 4, and the lowest (0.876) in module 2. We believe that one of the reasons for differences between module 2 and module 4 results is the different workload levels, with four module-level objectives addressed by module 3 and seven by module 4.

As part of the module-level pre-test and post-test surveys, 70% or more students reported feeling confident or very confident in their objective-related skills and knowledge for 21 (91.3%) of module-level objectives. For 12 (57.1%) of module-level objectives, 81% or more students reported this confidence level. This exceeds our expectations of 70%. However, analysis identified the need for course revisions to ensure improvement in meeting two module-level objectives, where the level of confidence was 52% and 54%. One is the ability to identify digital content management (DCM) tools that could be used for organizing language archives and DCM selection considerations. Another is the ability to determine DLA funding sources and budgeting principles. We attribute this finding to the fact that while these learning objectives are covered in required readings (presentations and examples), they were not reinforced through practical assignment tasks. Yet, even for objectives with lower-than-expected student confidence level demonstrated in the post-test, the difference between pre-test and post-test confidence level was significant.

The end-of-semester project-team-administered survey data indicated that between 91% (1 objective) and 100% students (4 objectives) developed confidence in their knowledge and skills related to course-level learning objectives. This was higher than the 85% expected by the project team. However, unlike with module-level objectives, we did not collect the pre-test survey data for course-level objectives at the beginning of semester.

The university-administered survey results demonstrated high levels of satisfaction with the course content (very good or excellent in 79% of student responses), teaching effectiveness (90%), and the course overall (79%). At the same time, the perceived amount of time to succeed in this course was reported to be higher than average by 63% of students, with 26% considering it much higher. We analysed the answers to open-ended questions, compared to the results of project-team-administered surveys, and used in course revision as the source of additional insight. For example, students suggested adding at least one live online class meeting. Student feedback indicates that a 5-week semester is not ideal for offering a course such as this one, with the variety and complexity of included content. This finding was expected because we developed the course with longer semesters in mind and assigned no required prerequisites, which resulted in a mix of relevant expertise levels among students in the class, with the majority (over 80%) being completely new to DLA topics and needing more time to develop confidence.

Community training and collaboration

This project, implemented by the interdisciplinary team of information scientists and linguists, also has the goal of collaborating directly with language communities to offer training on language archiving and obtain community feedback. The training consists of short workshops in which participants learn about the history, purpose, principles, and practicalities of digital archiving, and take part in hands-on activities designed to familiarize them with digital archives and provide experience with the process of adding an item to a DLA. While community workshops are still in progress, the planning of these collaborations in and of itself has introduced multiple insights that are relevant for future researchers prioritizing community-oriented participatory methods. While the needs of a language community differ in each community, we identify four areas that should be considered in the development of workshop materials for language communities: digital literacy, inter-generational collaboration, technology availability, and language accessibility.

Multiple factors such as socioeconomic status, proximity to urban areas, and community perception of technology, affect the acquisition of digital literacy – skills required to use digital

technologies – in communities. Language documentation projects have noted the lack of digital literacy as a barrier for community members to participate in digital resource creation (Frawley, Larkin, & Smith, 2017; Hausknecht et al., 2021; Mwanza, Molepo, & Goduka, 2018). This lack severely impacts the ability of a language community member to contribute to the creation of material in their language through digital archiving, which is detrimental to the language preservation (Domeji et al., 2019). When contributing to a DLA (by submitting an item or a collection), a depositor needs a level of digital literacy that allows performing the associated actions, such as uploading content from a separate device to a computer, file management, creating metadata, and filling out intake information on a web page.

In scenarios where community members lack sufficient digital literacy, DLA community workshops need to cover foundational digital literacy. However, when the digital literacy gap is prominent among older community members, but less pronounced in younger generations, inter-generational collaboration can alleviate the necessity of digital literacy training while simultaneously strengthening relationships between community members. One workshop in our project's series is designed for members of a language community that are university students. While the students have varying levels of proficiency in the language, partnering with older speakers in the community is mutually beneficial as they can support the older speakers navigate the technology used to contribute to archives and learn more about their language and culture from working with the older speaker.

The technology availability affects what happens both during and after the DLA community workshop. The workshops include hands-on activities, such as exploring a digital archive and preparing a file for deposit. To facilitate these activities, the participants must use a computer, that many of them don't have, so it is important to schedule the community workshop in a computer lab or bring laptops for the participants to use. In the non-university settings, access to a computer lab is fairly simple, but in other settings, this can present a substantial coordination challenge. It is also essential to note the technology the participant will be using after the workshop, including but not limited to the general access to computational technology as other factors can cause confusion: for example, differences between the operating systems or text / audio-visual editing software available for participants after the workshop versus what community participants used in training.

Building on ideas of general accessibility, the language used to present DLA to language community members is influential in facilitating understanding and garnering interest in digital archiving. For this reason, the way the content is presented to information science students or linguistics students must differ from the way it is presented to language community members. While the content remains largely unchanged, field-specific terminology should be modified to be intelligible to people without information science or linguistics background. For example, mentioning that the ability to query the metadata associated with items is one of the major advantages of DLA is not a persuasive point when the person has not been exposed to the concepts of metadata and querying. Similarly, expressing the importance of file naming conventions and organization of files must be substantiated in its practical utility: saving time and not losing things. The language used to discuss DLA must be accessible, both in the terminology that it employs and in the practical grounding of concepts.

Conclusion and next steps

The first offering of the project-developed LAM graduate course demonstrates overall success in curriculum development, as well as substantial demand among current students for coursework that is designed to develop the knowledge and skills necessary for successful stewardship of DLA, including community archives. Now that the course is officially added to UNT graduate catalogue as INFO5385, and students who completed it are sharing their experiences with fellow students, the awareness grows among our LAM student population. With 14 additional IMLS-sponsored

tuition stipends for the course, another class of 25 is expected to take INFO 5385 after revisions based on results of student evaluations completed so far and on the upcoming results of language community evaluation of project-developed learning materials.

For the longer (10-week) semester of 2025, we will add to INFO5385 the content covering two course-level objectives that did not reach the 70% student confidence target, break four instructor presentations into 2-3 shorter videos per module, add two live online meetings to the asynchronous course. The changes also include augmenting the data collection by adding the pre-semester survey for course-level objectives. The student feedback will be comparatively analysed and used in further refinements of learning materials.

Our team is also working on integrating relevant content from project-developed course modules into other UNT LAM courses and Indiana University Bloomington linguistics courses. We are disseminating preliminary results to obtain feedback from LAM and Linguistics educators and practitioners through presenting at professional meetings (Coronado & Zavalina, 2024; Zavalina & Paterson, 2024, etc.) and raise awareness of the project's products and opportunities it provides among the library and archival communities of practice, as well as among LAM faculty and students. Once the project is completed, project-developed learning materials will be made available as open access resources and are expected to make a solid contribution to addressing the needs of language communities through the extension of LAM professionals' preparation to stewarding DLAs.

Acknowledgements

The project presented in this paper is funded by the United States Institute of Museum and Library Services (grant RE-254860-OLS-23).

About the authors

Oksana L. Zavalina is Professor in Department of Information Science, University of North Texas, USA. She received her Ph.D. from University of Illinois, and her research interests are in the information organization and user needs in libraries and other repositories, including digital language archives. Dr. Zavalina can be contacted at Oksana.Zavalina@unt.edu

Alexandra O'Neil is Ph.D. student in Department of Linguistics, Indiana University Bloomington, USA. Her research interests are in the enhancing access to internet resources and technology for speakers of under-resourced languages. She can be contacted at aconeil@iu.edu

Shobhana L. Chelliah is Professor in Department of Linguistics, Indiana University Bloomington, USA. She received her Ph.D. from University of Texas in Austin, and her research interests are documentary linguistics and language archiving. Dr. Chelliah can be contacted at schellia@iu.edu

References

Al Smadi, D., Barnes, S., Blair, M., Chong, M., Cole-Jett, R., Davis, A., Hardisty, S., Hooker, J., Jackson, C., Kennedy, T., et al. (2016). Exploratory user research for CORSAL. https://digital.library.unt.edu/ark:/67531/metadc1707416/

Aljalahmah, S., & Zavalina, O.L. (2023). Exploration of metadata practices in digital collections of archives with Arabian language materials. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114295/

Bird, S., & Simons, G. (2003), Seven dimensions of portability for language documentation and description. https://doi.org/10.1353/lan.2003.0149

Burke, M. (2021). Collaborating with language community members to enrich ethnographic descriptions in a language archive. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2021). https://doi.org/10.12794/langarc1851172

Burke, M. (2023). Designing Archival Collections to Support Language Revitalization: Case Study of the Boro Language Resource. Thesis or dissertation, University of North Texas. https://doi.org/10.12794/metadc2137570

Burke, M., Zavalina, O.L., Chelliah, S.L., & Phillips, M.E. (2022). User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users. Language Documentation and Conservation, 16, 1–24. http://hdl.handle.net/10125/74669

Chelliah, S.L. (2021). Why language documentation matters. Springer

Chelliah, S.L. (2023). Making photographs in language archives maximally useful: Metadata guidelines for community and academic depositors. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114301/

Cook, T. (2013). Evidence, memory, identity, and community: four shifting archival paradigms. Archival Science, 13(2), 95–120. https://doi.org/10.1007/s10502-012-9180-7

Coronado, S.I., & Zavalina, O.L. (2024). Digital language archiving: Reuse and adaptation of non-LIS learning materials in LIS education. In, The Ethics and Evolution of Truth and Information: Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2024. https://doi.org/10.21900/j.alise.2024.1750

Dale, M., Basumatary, P., Iqbal, J., Khullar, R., & Shaikh, M. (2022). Case study of using Facebook groups to connect community users to archived CoRSAL content. Language Documentation & Conservation, 16, 399–416. https://hdl.handle.net/10125/74685

Domeij, R., Karlsson, O., Moshagen, S., & Trosterud, T. (2019). Enhancing Information Accessibility and Digital Literacy for Minorities Using Language Technology– the Example of Sámi and Other National Minority Languages in Sweden. In, Perspectives on Indigenous writing and literacies (pp. 113–137). https://doi.org/10.1163/9789004298507_007

Frawley, J., Larkin, S., & Smith, J.A. (2017). Indigenous Pathways and Transitions into Higher Education: An Introduction, p. 3–11. Springer, Singapore. https://doi.org/10.1007/978-98110-4062-7_1

Genovese, T.R. (2016). Decolonizing archival methodology: Combating hegemony and moving towards a collaborative archival environment. AlterNative: An International Journal of Indigenous Peoples, 12(1), 32–42. https://doi.org/10.20507/AlterNative.2016.12.1.3

Harris, A., Gagau, S., Kell, J., Thieberger, N., & Ward, N. (2019). Making meaning of historical Papua New Guinea recordings: Collaborations of speaker communities and the archive. International Journal of Digital Curation, 14(11), 136–149. https://doi.org/10.2218/ijdc.v14i1.598

Hausknecht, S., Freeman, S., Martin, J., Nash, C., & Skinner, K. (2021). Sharing indigenous knowledge through intergenerational digital storytelling: Design of a workshop engaging elders and youth. Educational Gerontology, 47(7), 285–296. https://doi.org/10.1080/03601277.2021.1927484

Henke, R.E., & Berez-Kroeker, A.L. (2016). A brief history of archiving in language documentation,

with an annotated bibliography. Language Documentation & Conservation, 10, 411–457. http://hdl.handle.net/10125/24714

Hinton, L. (2001). The Use of Linguistic Archives in Language Revitalization: The Native California Language Restoration Workshop, p. 419–428. Brill, Leiden, The Netherlands https://doi.org/10.1163/9789004261723_033

Jones, M. (2019). Collections in the expanded field: Relationality and the provenance of artefacts and archives. Heritage, 2(1), 884–897. https://doi.org/10.3390/heritage2010059

Mwanza, A.J., Molepo, J.M., & Goduka, N. (2018). Addressing the digital literacy gap in the use of information and communication technologies in the dissemination of indigenous knowledge: a rural community experience. Indilinga: African Journal of Indigenous Knowledge Systems, 17(22), 154–166. https://www.ajol.info/index.php/indilinga/article/view/186128

Olson, E.A. (2023). Mrs. his name: reparative description as a tool for cultural sensitivity and discoverability. Journal of Western Archives, 14(1), Article 10. https://digitalcommons.usu.edu/westernarchives/vol14/iss1/10/

Paterson, H.J., III (2023). OLAC and serials: An appraisal. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114296/

Pugh, M., & Christen, K. (2011). Opening archives: Respectful repatriation. The American Archivist, 74(1), 185–210. https://doi.org/10.17723/aarc.74.1.4233nv6nv6428521

Roeschley, A., & Kim, J. (2019). 'Something that feels like a community': The role of personal stories in building community-based participatory archives. Archival Science, 19(1), 27–49. https://doi.org/10.1007/s10502-019-09302-2

Rolan, G. (2017). Agency in the archive: a model for participatory recordkeeping. Archival Science 17(3), 195–225. https://doi.org/10.1007/s10502-016-9267-7

Wasson, C., Holton, G., & Roth, H.S. (2016). Bringing user-centred design to the field of language archives. Language Documentation & Conservation, 10, 641–671. http://hdl.handle.net/10125/24721

Zavalin, V. (2023). Ukrainian archival metadata in WorldCat: Exploratory analysis. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114298

Zavalina, O.L. (2023). Language archiving training: A case study of a metadata course in Library and Information Science graduate program, 2020 - 2023. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114298

Zavalina, O.L., & Chelliah, S.L. (2021). Exploring language archiving education for information professionals and interdisciplinary collaboration to support information access. Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2021. https://www.ideals.illinois.edu/items/118795

Zavalina, O.L., & Paterson, H.J., III (2024). Developing graduate curriculum for digital language archive stewardship. In, The Ethics and Evolution of Truth and Information: Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2024. https://doi.org/10.21900/j.alise.2024.1657

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

The inevitability of AI: a study of undergraduate students' perceptions of AI tools in their future careers

Mónica Colón-Aguirre and Kawanna Bright

DOI: https://doi.org/10.47989/ir30iConf47308

Abstract

Introduction. Artificial intelligence (AI) tools garner more attention every day, questions have arisen regarding their possible negative impacts in future job markets. Some predict a potential for massive job losses, especially in high-skilled jobs. This study seeks to explore undergraduate students' perceptions of how these tools might affect their future careers.

Method. This study follows a case study design, employing phenomenological interviews as a research method. The data set was made up of interviews with 17 undergraduate students.

Analysis. Data were analysed by employing constant comparative analysis, with various rounds of coding including the creation of open, axial, and structural codes.

Results. Students saw AI tools as an inevitable part of their future work. Participants expressed their intention to learn how to optimize their use of various tools, which they see as having the potential to positively benefit them in their future careers. They do not perceive AI to be a viable substitute for their skills, especially in terms of identifying misinformation and emotions.

Conclusions. Academic institutions must provide curricular spaces which allow students to harness the power of AI tools. While employers should also make efforts to train employees to make the most of AI tools.

Introduction

A recent article published in *Fortune* described the advent of AI technologies' effect on the tech industry as '*corporate Ozempic*' (Confino, 2024). This metaphor refers to the diabetes drug's secondary use as a weight loss drug which allows those who use it to shed weight without admitting its use. According to Scott Galloway, the rash of tech sector layoffs were due to companies using AI tools which reduced the need for certain positions and made many tech employees' work redundant (Confino, 2024). This meant that by early 2024 around 26,000 jobs in the usually high paying, technology sector had vanished (Allyn, 2024).

What made these layoffs even more alarming than their sheer magnitude was that they affected workers in corporate giants like Meta, Amazon, Microsoft, Google, TikTok and Salesforce and were almost exclusively high skilled workers (Allyn, 2024). Although the initial causes of these layoffs have been attributed as an effort by organizations to return to pre-COVID pandemic numbers after a rush of pandemic hiring (Allyn, 2024) there are some, like Galloway, who do not find it coincidental that the massive layoffs in what used to be considered a '*safe*' industry coincided with the advent of generative AI tools such as ChatGPT-4.

Abbreviated literature review

The issue of the effects of automation, specifically AI systems, reshaping the nature of work and causing mass job losses is not new (Duchessi et al., 1993). However, the topic has been revisited in recent years after the unveiling of ChatGPT-4 and advances in related technologies such as cloud computing, virtualization, and the Internet of Things (Morandini, 2023). With the renewed interest in AI and its possible effects on the global economy come questions regarding how these technologies will affect specific employment sectors. The International Monetary Fund (IMF) predicts that up to 40% of jobs around the world will be affected in one way or another by AI, especially high-skilled jobs (Georgieva, 2024). The effects of AI technologies in the workplace are expected to be felt within the next few years, causing disruption in many workers' lives with those who can successfully harness AI applications seeing increases in their wages and productivity, while those who can't may be out of the job market completely (Georgieva, 2024).

In the near future it is likely that many organizations will be adapting new operation models which will require a reconsideration of employees' skills as the effects of AI automation will require the development of new skills and the further development of others (Morandini et al., 2023). There appears to be a generalized agreement that so-called '*soft skills*,' such as critical thinking, problem-solving, and self-management will become more important as workers incorporate AI systems into their day-to-day work (Morandini et al., 2023; World Economic Forum, 2020).

In terms of the type of occupations most affected by AI automation, some predict that it will be those employing high skilled workers, including sectors such as postsecondary education in areas like English, foreign language, history, law, and social sciences (Felten, 2023). This anticipated AI-induced destabilization is expected to affect higher education, as institutions of higher learning are both the main trainers of highly skilled workers and employers of highly skilled workers.

Explorations relating to the early adaptation of AI tools in higher education have found that the best strategy for higher education is to adapt these tools into the academic context rather than fighting against them (Neumann et al., 2023; Rasul et al., 2023). Many of these early studies have focused on providing faculty and administrators in higher education guidance into how to best incorporate and adapt to these tools.

A rising number of works have begun to explore students' experiences and perceptions of AI. Most of these studies have focused on students' adoption of AI for use in higher education. Results of these studies have been similar, with most showing students have a general awareness of AI (Dabirian & Swarat, 2024), but relatively low to mid-level use of AI for academic purposes (Chen et

al., 2024; Dabirian & Swarat, 2024; Marshik et al., 2024), though they do see value in AI for academic purposes (Al-Abdullatif & Alsubaie, 2024; Chen et al., 2024). What has not been explored is students' perceptions of the impact of AI on their future careers and work, especially the perception of those who are at the cusp of their academic career and going into the job market in the near future.

Following the iConference theme, '*living in an AI-gorithmic world*,' this short paper explores undergraduate students' perceptions of generative AI tools in the context of higher education as it pertains to their expectations of its influence on their future careers. In this research, the perspective of college students about to enter the new job market which has been heavily influenced by AI tools is explored. This population is important since it allows us to further our understanding of the perceived challenges and advantages these technologies will present to new employees who are entering the job market immediately after what some are referring to as the advent of the 4th industrial revolution, or industry 4.0. (IBM, n.d.). This study will offer the opportunity to explore the ways in which students expect AI technologies to affect their prospective careers and which strategies they are incorporating to help position themselves favourably in the upcoming job market. The main research question driving this project is: What are students' perceptions of how AI tools will influence their future careers?

Method

Research design

This research is part of a larger scale research project which uses an embedded case study design. As presented by Robert Yin (2018), an embedded case study is a type of single case study in which a single organization serves as the primary unit of analysis; however, data is gathered from various elements, or 'embedded' units of analysis, within that organization. In this case The University is the primary unit of analysis. Although data is being gathered from various populations at The University, including faculty and librarians, this work focuses on the findings related to undergraduate students. The University is the flagship campus of a system in the southern region of the United States and classified by the Carnegie Classification of Institutions of Higher Education as a public, very high research activity university.

The University has not prohibited the use of AI tools for academic purposes but has warned students that its use in coursework is left to their instructor's discretion, and that the unauthorized use of such technologies may be classified as cheating according to the institution's honour code. Data for this study are being gathered through one-on-one interviews. The question protocol used with students gathered information regarding their general educational experiences, experience with AI tools, and their perceptions regarding applications of these tools to various professional scenarios, including future career and occupational expectations.

Sampling

Permission was obtained from the institutional IRB of the lead researcher prior to the start of data collection. Students were recruited through the candidate pool available through one of the academic units and offered the opportunity to join a raffle for one of three $50-dollar online retailer gift cards as a participation incentive. A total of 17 students participated in interviews conducted using an online video call platform. The interviews produced eight hours and 20 minutes of recordings which were captured digitally on the platform, which was also used to produce automatically generated transcripts. Transcripts were reviewed by the lead researcher to ensure accuracy. All participants were assigned pseudonyms which were presented and approved at the time their interview took place.

Description of the study sample

The use of a recruitment tool meant that the students in the sample were all affiliated with one of the colleges at The University. This college specializes in areas of communication and information studies. The areas are closely related to the socio behavioral sciences and represent to occupations which depend on creating close social connections to clients and audiences, as well as a great deal of creativity, with the goal of advancing social causes by understanding and explaining behavioral and social phenomena (Gerstein, et al., 1988).

Participants' average age was 20 years old. Ten of the participants were seniors, while four were juniors, and three were sophomores. There were no freshmen in the sample. When asked to identify themselves, 12 identified as female, three as male, while two did not specify any gender.

Data analysis

The data were analysed inductively and went through several rounds of coding, starting with open coding which allowed for the identification of initial codes. Several rounds of axial coding followed, allowing for the collapsing of several codes to form more abstract code categories. Further coding was performed using structural coding, a technique that focuses on coding and creating initial categories on a data corpus which allows for comparisons between specific segments to be established in relation to their commonalities, differences, and relationships (Saldaña, 2016). This type of coding is appropriate for qualitative data sets employing multiple participants and where semi structured instruments are used (Saldaña, 2016). All coding categories used in the analysis were chosen by mutual agreement between the researchers.

This research methodology allowed the researchers to ascertain the contextual nature of the participants' awareness, use, and perceived application and value of AI tools. The findings presented here stem from the preliminary categories created for the specific question related to the participants' perception of the influence that AI tools will have in their careers.

Results

The general perception among participants in this study was that AI tools will change or deeply affect their future careers and the occupations which they are planning to enter upon graduation. Overall, students based their value judgments regarding AI tools and their possible professional uses on the notion of the inevitability of AI. This term expresses how most students saw AI as a new fact of life, which is shaping their academic experiences and will shape their professional experiences into the future. Due to the perception among students that AI is the new sine qua non of the workplace, most expressed the need to work with and master AI tools to improve their future career prospects.

For example, Morgan, a senior with a mass communication concentration, indicated that AI tools will be ubiquitous in their field, therefore the imperative will be to learn how to use it: '*I feel like it's a matter of working with ChatGPT and not fighting it*', '*cause inevitably everyone's gonna use it…*' This sentiment was echoed by Gerry, a senior majoring in advertising, who saw AI as a helpful tool they plan to use in their future career:

> *Like I said, time efficiency, so it'll like help me get stuff done quicker. I also think it could add to like my portfolio as I get better at using it…but I think it will make it [their future job] easier, it'll make most things easier.*

While many saw AI as the latest technological change to which they must adapt if they want to be competitive in the workplace, others saw it as a threat - one they must learn to grapple with. For example, Ozzie, a senior mass communication major, explained:

> *I definitely think I've heard a lot of buzz about certain jobs being impacted because of the abilities of AI. Like we don't necessarily need as many people doing certain things. I think in my career, if anything, it'll be a tool just for writing quick documents or creating presentations faster or better.*

This same feeling was expressed by Pax, a junior public relations major:

> *I think it'll get rid of a lot of jobs just because like a lot of these things, they don't need to pay people to do it. If you can just have a computer, do it. But I think also like, it'll make it more convenient, and I mean, I think that's really it. Like it'll just, it'll get rid of a lot of jobs, especially in the field that I wanna go into because a lot of that can just be done by a computer.*

Participants in the study also expressed concerns about the advent of AI in the professional arena, especially in terms of fear for what will happen if AI tools go unchecked. Some of the most pressing areas of concern related to the capacity of spreading and generating misinformation using AI, and the perception that AI has the potential to dehumanize human communication. The students interviewed shared that these perceived AI shortcomings are the areas where they have the most to contribute to the professional arena. As Clarice, a sophomore mass communication major, expressed:

> *like we're gonna have a lot of things like quantity wise, but not necessarily quality. We were not actually checking to see if these things are, you know, necessarily relevant or correct. And so that kind of scares me as to what that looks like, especially in my realm of us putting out information where we have to be quick and like inform people about different things that are going on in the world. And you would hate to like inform somebody and create misinformation and then it leads to some type of hysteria or negative impact.*

Similarly, Nia, a junior information science major, shared her fear that AI will replace people in jobs and how the perceived weakness of AI's lack of emotion can serve as the competitive edge she will have when she joins the workforce:

> *so, as people do say that AI is gonna take over and especially with my major in tech, I feel like people are most scared about that. But I feel like there's just some things that AI, like I said, lacks that emotional connect[ion] with information science. That's something that we like emphasize a lot is learning the soft skills and yeah. Privacy. And I feel like AI...just cannot understand that.*

Despite a generalized feeling that AI is not a replacement to humans in the work environment due to its inability to understand and react to human emotions, the need to harness AI's power during their college preparation is still in the participants' minds. Some even expressed how they are thinking of adapting or reconsidering the courses they take or their college major selections and their current programs of study to help position themselves favorably in the future job market. This is exemplified by Iggy, a sophomore information science major:

> *...I wanna be like a data scientist, or 'cause in my job search I have seen a lot of emerging jobs, for like AI scientists and AI specialist. And see that's sort of like the thing I'm stuck at right now. Like, I don't know, should I be focusing on AI? Or like, yeah, it's big, it keeps growing, but...10 years from now, is it all gonna be illegal or regulated in some way? And the same with the data scientist route. I go down that, within five years from now, are they just gonna have robots doing all that? So, like, yeah, I'm kind of stuck at a crossroads.*

Conclusions

As demonstrated by the findings in this study, participants want to acquire and develop skills and knowledge which will help them feel safe from being replaced by AI tools in their chosen careers. The general sentiment expressed by study participants is that AI is an inevitable aspect of their future job prospects. This leads the participants to take an attitude more accommodating towards learning how to best use AI tools and incorporate them into their current course of study in a way that best positions them with the skills needed in the job market. This attitude is accompanied by the generalized belief among participants that AI is not a panacea or infallible, leaving room for their skills, as human actors, to best position themselves in the future job market. This is reflected in the attitude that AI tools can't do everything that humans can, including identifying misinformation, and providing the emotional connection that their socio-behavioral disciplines require.

In positioning themselves as humans with better capabilities for identify misinformation, human emotions, and being able to take specific actions to curtail these shortcomings, participants expressed their confidence in terms of being competitive in the future job market. It is important to highlight here that they do see the advantages of employing AI in their future careers but see it as yet another technological tool which can help them become better at their jobs, and not as a replacement for their skills and human essence in the future occupations.

In general, the students in this sample perceived AI tools as useful additions to their professional toolkit, but not as a replacement for their talents as employees or even as students. This finding has implications for multiple stakeholders in academia and beyond. For academic institutions, it highlights the importance of creating a curriculum which adopts AI tools instead of shunning or prohibiting it, as has already been highlighted in the literature (Neumann et al., 2023; Rasul et al., 2023). Adapting instead of resisting new AI technologies will help bring about a new generation of technologically literate, well-rounded professionals who can harness the power of AI and future technologies to optimize the work they do across all industrial sectors. But the onus of providing training and learning to adapt to the changes brought on by AI do not fall squarely on the shoulders of academia. It is also essential for employers across industries to help train and develop their employees so that they can utilize this and any upcoming technologies to their full potential. This has the potential to benefit not only organizations but also society.

About the authors

Mónica Colón-Aguirre is Assistant Professor at the University of South Carolina's School of Information Science, USA. She holds a PhD from the University of Tennessee, Knoxville. Her research includes topics related to information science, academic libraries, and LIS education.

Kawanna Bright is Assistant Professor of Library Science at East Carolina University, USA. She holds a PhD in research methods and statistics from the University of Denver. Her current research focuses on assessment in libraries; equity, diversity, and inclusion (EDI) in libraries and the liaison role in academic libraries.

References

Al-Abdullatif, A. M., & Alsubaie, M. A. (2024). ChatGPT in learning: Assessing students' use intentions through the lens of perceived value and the influence of AI literacy. Behavioral Sciences, 14, 845. https://doi.org/10.3390/bs14090845

Allyn, B. (2024, January 28). Nearly 25,000 tech workers were laid off in the first weeks of 2024. Why is that? NPR. Retrieved from: https://www.npr.org/2024/01/28/1227326215/nearly-25-000-tech-workers-laid-off-in-the-first-weeks-of-2024-whats-going-on

Chen, K., Tallant, A. C., & Selig, I. (2024). Exploring generative AI literacy in higher education: Student adoption, interaction, evaluation and ethical perceptions. Information and Learning Sciences. http://dx.doi.org/10.1108/ILS-10-2023-0160

Confino, P. (2024, February 29). All the tech layoffs are because AI is like 'corporate Ozempic'—it trims the fat and you keep the fact you're using it a secret, says marketing guru Scott Galloway. *Fortune*. Retrieved from: https://fortune.com/2024/02/29/tech-layoffs-ai-corporate-ozempic-meta-apple-salesforce-google-scott-galloway/

Dabirian, A., & Swarat, S. (2024). Artificial intelligence in higher education: Community perceptions at a large U.S. university. IT Professional, 26(4), 92-96.

Duchessi, P., O'Keefe, R., & O'Leary, D. (1993). A research perspective: Artificial intelligence, management and organizations. *Intelligent Systems in Accounting, Finance and Management*, 2(3), 151-159.

Felten, E., Raj, M., & Seamans, R. (2023). How will language modelers like ChatGPT affect occupations and industries?. *arXiv preprint* arXiv:2303.01157.

Gerstein, D. R., Luce, R. D., Smelser, N. J. & Sperlich, S. (1988). The Behavioral and Social Sciences: Achievements and Opportunities. National Academy Press (Committee on Basic Research in the Behavioral and Social Sciences): Washington, D.C.

Georgieva, K. (2024, January 14). AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity. *IMF Blog*. https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity

IBM (n.d.). What is Industry 4.0? Retrieved from: https://www.ibm.com/topics/industry-4-0#:~:text=We%20are%20now%20in%20the,productively%20across%20the%20value%20chain.

Neumann, M., Rauschenberger, M., & Schön, E. M. (2023, May). "We Need To Talk About ChatGPT": The Future of AI and Higher Education. In *2023 IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG)* (pp. 29-32). IEEE.

Marshik, T., McCracken, C., Kopp, B., & O'Marrah, M. (2024). Student and instructor perceptions and uses of artificial intelligence in higher education. Teaching of Psychology. https://doi.org/10.1177/00986283241299745

Morandini, S., Fraboni, F., De Angelis, M., Puzzo, G., Giusino, D., & Pietrantoni, L. (2023). The impact of artificial intelligence on workers' skills: Upskilling and reskilling in organisations. *Informing Science*, 26, 39-68.

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A. & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching*, 6(1).

Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). Sage.

World Economic Forum. (2019). *World Economic Forum Annual Meeting* [Conference session]. https://www.wefo-rum.org/events/world-economic-forum-annual-meeting-202

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th edition). Sage.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

The smart archive management practicing pipeline: a virtual online learning platform for AI literacy development in archival science

Lanxi Dong, Shuning Tang, Yuan Cheng, and Ping Wang

DOI: https://doi.org/10.47989/ir30iConf47305

Abstract

Introduction. In order to develop AI literacy for Chinese students majoring in archival science, this study presents the design and implementation of the Smart Archival Learning Platform (SAMPP), providing an immersive environment to understand AI-driven archival workflows, develop AI skills, and improve AI competence.

Method. The implementation of the SAMPP follows a three-step strategy: 1) an introductive and immersive 3D virtual environment to fix students' attentions on AI learning; 2) AI-powered archive management tasks for skill development, and 3) an automated online exhibition project to enhance creativity and competency.

Analysis. As an impact evaluation, both questionnaires and interviews are conducted with 55 undergraduate students majoring in archival science at Wuhan University before and after using the SAMPP. The results of the questionnaires are used for sentiment analysis, and students' attitudes toward the platform are explored through interviews.

Results. Before using the SAMPP, 66% of students had negative perceptions of AI learning. Afterwards, 58% reported positive attitudes, showing enhanced confidence in using AI for archive management. Interviews reinforced these findings.

Conclusion. The SAMPP enhances AI literacy and creativity among archival science students, making AI learning accessible and engaging. Future improvements should integrate advanced AI models to expand its capabilities.

Introduction

Archive management can be defined as a process that pertains to the informational value of archives efficiently and systematically (Standardization, 2016) (Ilerbaig, 2019). When artificial intelligence (AI) technologies are rapidly advancing and are impacting all areas of modern industries worldwide (Black et al., 2024), archive management will not be spared from such a fierce technological revolution. While AI technologies empower archive management with unparalleled efficiency and resources, a question that arises is how an archivist defines him/herself in such an era. Is he/she just a skilful practitioner freed from daily archive management routines with the assistance of AI technologies, or a contributor for developing domain specific AI products for archive management, or an empowered storyteller who can produce creative archive-storytelling works using multiple AI models (e.g., pre-trained large language models)?

AI is now at the top of the agenda for many educators who are educating the next generation worldwide. It is undeniable that the more AI-literate a person is, the easier he/she can choose a way to collaborate with AI. As a positive response and with a pilot study, a virtual online learning platform, named the Smart Archive Management Practicing Pipeline (SAMPP), was designed, and developed by educators from the department of archive management and government information of Wuhan University. The platform was envisioned to become a venue for undergraduate students majoring in archival science to gain a comprehensive understanding of how AI is transforming the workflow of archive management.

This paper presents the details involving the design and implementation of the Smart Archive Management Practicing Pipeline, an online learning platform designed and offered by Wuhan University for LIS students, especially those majoring in archival science. The ensuing narrative establishes a brief survey on virtual online learning platforms for educational purposes and gives pertinent information, including the design strategies for progressive AI literacy building, system components, and student feedback based on sentiment analysis.

Related works

Virtual simulation platforms have been widely employed in various disciplines, particularly in fields such as computer science, medicine, and engineering. For example, in dentistry, a virtual simulation platform for teaching pulpotomy has been shown to significantly enhance the effectiveness of educational practices(Lu et al., 2022). In engineering education, the Mars Exploration Control Virtual Simulation Experiment Platform (MEC-vslab) was designed to address the challenge of limited integration between control engineering theory and practice (Wang et al., 2024). Similarly, in the domain of physical education, virtual simulation platforms utilizing optical principles have improved students' understanding of complex concepts(Zhang, 2024). These platforms are valued for their ability to reduce experimental costs, improve efficiency, and enhance safety by simulating real-world scenarios. Furthermore, advancements in technology have enabled real-time monitoring and personalized teaching through these platforms.

The application of AI technology in virtual simulation has also proven to be highly effective in a range of fields. For instance, AI-based virtual robots have been employed to improve simulation efficiency and accuracy, providing valuable data support for enterprise decision-making (Wenjing, 2025). In educational contexts, the integration of AI technology has led to more immersive and realistic simulations, enhancing learning outcomes. However, most of these platforms focus primarily on improving platform performance and creating more engaging experimental environments, rather than on enhancing users' understanding of AI or fostering the development of AI skills, particularly in social sciences.

The primary objective of the Smart Archive Learning Platform (SAMPP) is to address this gap by improving the AI literacy of students with a liberal arts background, specifically those studying archival science. SAMPP seeks to introduce these students to AI technologies and familiarize them

with their application within the field of archive management. By doing so, it offers a blueprint for the development of similar platforms in other disciplines that traditionally lack a focus on STEM education.

Design strategy

AI literacy is conceptualized as a holistic construct comprised of proficiency areas of knowledge, awareness, skills, competencies, and experience (Pinski & Benlian, 2024). The goal of SAMPP is to include the above mentioned five 'enabling constructs' as much as possible so students can better situate themselves in this era of AI. Inspired by the AI literacy developing strategy framework proposed by (Black et al., 2024), this section describes a three-step strategy to engage undergraduate students from archival science in AI learning progressively. The strategy provides a conceptual and structured framework for virtual AI learning platform development. Moreover, the strategy was also designed with the contemporary educational background in mind. In an investigation conducted by educators from the department of archival science and government information of Wuhan University, more than 60 percent of students investigated expressed their diffidence or anxiety on developing AI literacy.

The strategy aims to provide students with a full experience that runs in parallel with the lifecycle of record management so they are aware of the power of AI technologies, can acquire relevant knowledge, develop skills, and enhance competencies. The details are illustrated in Figure 1.

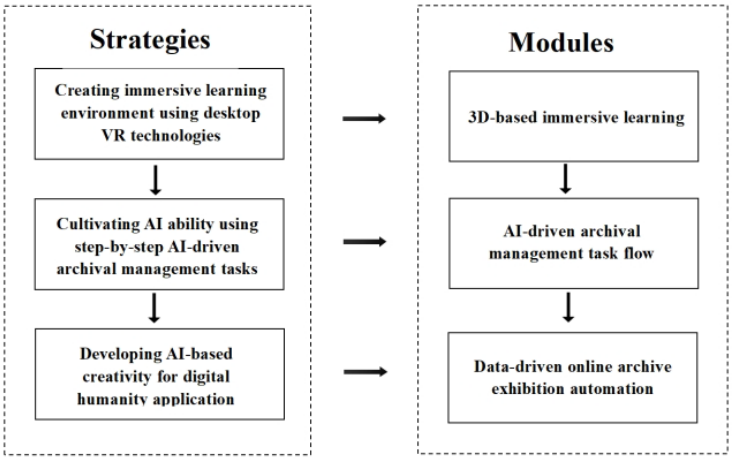


Figure 1. The correspondence between the 3-step AI literacy developing strategy and the SAMPP modules

Critical strategy1-Easing AI learning diffidence

As the first step of the progressive AI literacy developing strategy, the purpose of critical strategy 1 is not to rush the appearance absurd AI models in front of the beginners, but to relieve their unfamiliarity, anxiety, and diffidence when entering into a new world. Immersive learning methods have been shown to increase students' interests and commitments to the task at hand, particularly when combined with role-playing activities, which foster a sense of responsibility (Rowe et al., 2011). The sooner and more deeply that students immense themselves into the learning environment, the more eager and natural they feel to try learning about AI. A 3D desktop environment together with VR devices are on top of the list when building such an immersive environment. VR in training provides the capability of immersing learners in simulated environments that are digital twins of real-world environments, offering training scenarios in a controlled, risk-free setting (Bouloukakis et al., 2019) (Rossau et al., 2019). A typology of role-play learning designs would enable higher education teachers to select role-play approaches which

could be best used to provide learning opportunities to promote cognitive, affective and/or psychomotor domains of learning (Rao & Stupans, 2012).

Critical strategy2- Cultivate AI skills and accumulate AI knowledge

It is crucial for students to experience AI through active engagement in activities and projects(Black et al., 2024). Fundamental archive management involves a sequence of atomic tasks.

An atomic task refers to a set of activities that are executed as a whole and as an indivisible unit along the archive management pipeline, which includes decontamination, deviation correction of digital records, metadata registration, retention, and disposal. SAMPP is intrinsically an exploratory digital; environment that lets students comprehend the complex nature of AI and its application in archive management. By integrating AI algorithms into each of these tasks, students can be aware of how AI improves efficiency, accuracy, and decision-making in archive management workflow. The sequential nature of these tasks forces students to engage deeply with AI technologies, encouraging them to think critically about how general AI models can be fine-tuned for domain-specific applications. Besides enhancing the students' technical skills, students can acquire knowledge of AI, e.g., AI models, data for AI, and AI tools.

Critical Strategy3- Facilitate exploratory experiences and apply AI knowledge

The third level of this strategy focuses on introducing students into an exploring experience so they will have the opportunity to apply AI in the field of digital humanities. Archives represent a rich source of historical and cultural information, and this strategy aims to inspire students to think innovatively about how AI can help reveal and present the value of archival data to the public. Through this process, students are encouraged to consider questions such as, 'How can AI help showcase the historical significance of archives to a broader audience?' This final stage of the strategy not only deepens students' understanding of AI's potential in archive management but also enhances their ability to think creatively about the intersection of technology and the humanities.

System overview

The implementation of SAMPP incorporates both the exercise- and the artifact-based learning methods. Both methods are identified as formal learning methods that provide planned and structured learning activities (Marsick, 2001). The exercise-based learning methods use either traditional or interactive exercises to convey AI-specific content that relates to AI's learning abilities, while artifact-based learning utilizes AI as part of the learning method.

Here in SAMPP, the AI concepts/models/algorithms were recontextualized within the archive management process, making them easier to grasp. Critical steps in record management lifecycle are taken into considerations and serve as anchors, which include the creation, maintenance, use and disposition of records (Penn et al., 1994).

SAMPP consists of three modules that align with each phase of the strategy. It provides a structured pipeline of tasks, each of which expands upon and supports the strategic concepts outlined above. The three modules are as follows:

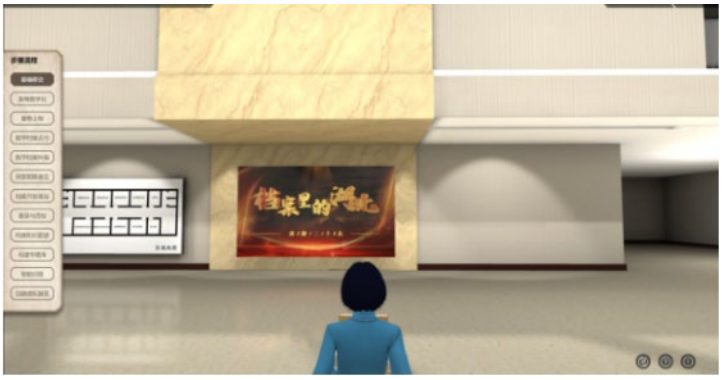
Module 1: An introductive 3D immersive scenery

In response to the critical strategy 1, a 3D scene is built at the beginning of the experimental procedure not only to create an immersive learning environment but as a simulation of the initial steps of archive management, namely records transfer and appraisal. According to Christian Hartmann et al.'s (Hartmann et al., 2023) experiment, immersive learning methods can increase the perception of presence, thereby obtaining more semantic knowledge. Real scenarios can also stimulate students' interest, inspiring them to explore autonomously, enhancing learning efficiency, and better promoting knowledge acquisition and understanding (Hartmann et al., 2023). At the same time, incorporating 3D technology at the beginning of the experimental procedure

enables students to engage with a more immersive manner, moving away from the abstraction of theoretical learning and concretizing management phases. The virtual reality stage stimulates students' confidence in learning and keeps them focused on the learning process.

In this module, a 3D replica of Hubei provincial archives is created using the 3Ds MAX software package. Immediately after logging in the SAMPP, a student is assigned with the role, who is on duty of transferring archive files to Hubei provincial archives, as it is illustrated in Figure.2-(a). This provides a first-person perspective allowing students to carry out the experiment at the very beginning. The role switching takes place when archive files are transferred between the girl in blue and the girl in white. The student is assigned to the role of the future archivist, in order to continue with the following experimental tasks, as it is illustrated in Figure.2-(b). Also in this 3D scene, we should change to let students focus on what a future archivist should do, as it is illustrated in Figure.2-(c), which enables a smooth transition from the immersive 3D environment to AI technique learning.

The highly realistic scenery enhances the visual attraction of the experiment, thereby focusing more on knowledge acquisition and improving their learning efficiency.



(a) A 3D replica of Hubei provincial archives



(b) Role switching, the girl in white acts as the future archivist



(c) A Shot cut of the 3D scene to AI-driven archive management task flow

Figure 2. Module 1-The 3D immersive scene at the beginning of the AI-learning procedure for students majoring in archival science.

Module 2: AI-driven archive management task flow

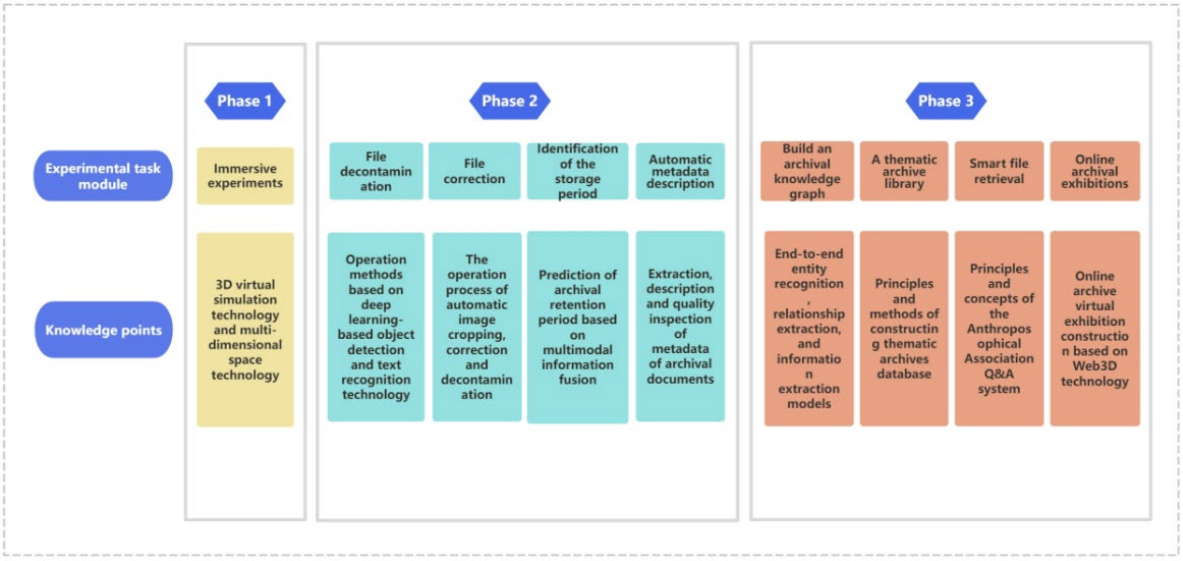


Figure 3. The archive management workflow of SAMPP

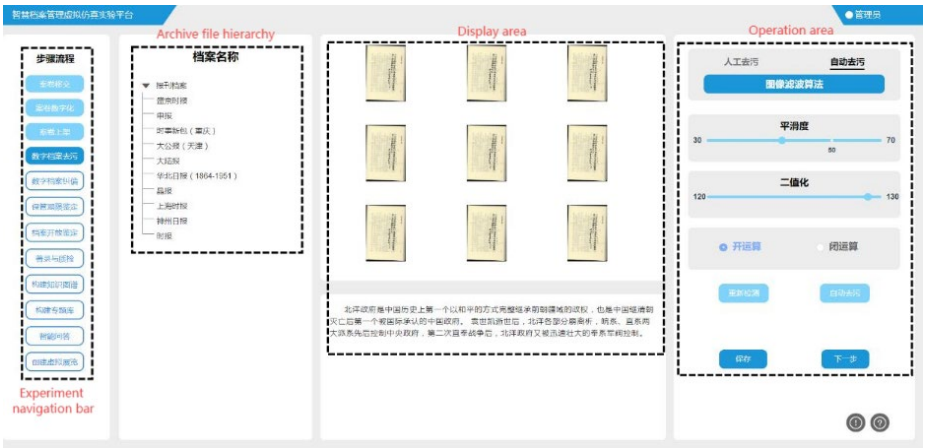
In response to the critical strategy 2, a task flow composed of archive management routine tasks is organized so students can have a preliminary comprehension over AI's impacts on archive management, as it is illustrated in Figure.3. The task flow begins with the intelligent image correction, and decontamination using computer vision algorithms, following by metadata registration based on object detection and text recognition, ending with the retention and disposal prediction based on information fusion. In each task, both manual and AI methods are presented for comparison. Combining algorithm principles with application results in the experimental process helps students fully understand each algorithm's workings and develop habits of applying technology to archival decontamination. Close interaction with AI technologies facilitates students' exploration of AI-enabled archive management, teaching them to carry out archive management activities with edgy technologies, developing their management skills, and establishing a knowledge base for AI applications. Screenshots of the module-2 is presented in Figure.5.

Intelligent Decontamination of Digital Archives: Two major steps are presented: manual decontamination and correction of individual documents and intelligent decontamination and correction. This allows students to experience the differences between manual and AI methods in digital archive management in terms of efficiency and quality. Students can choose from AI processing algorithms, such as the projection profile method, the RGB processing, and the integer Bresenham algorithm, for students to choose from. By matching algorithm types and adjusting parameters to achieve optimal results, students can more intuitively experience the different effects of various technologies.

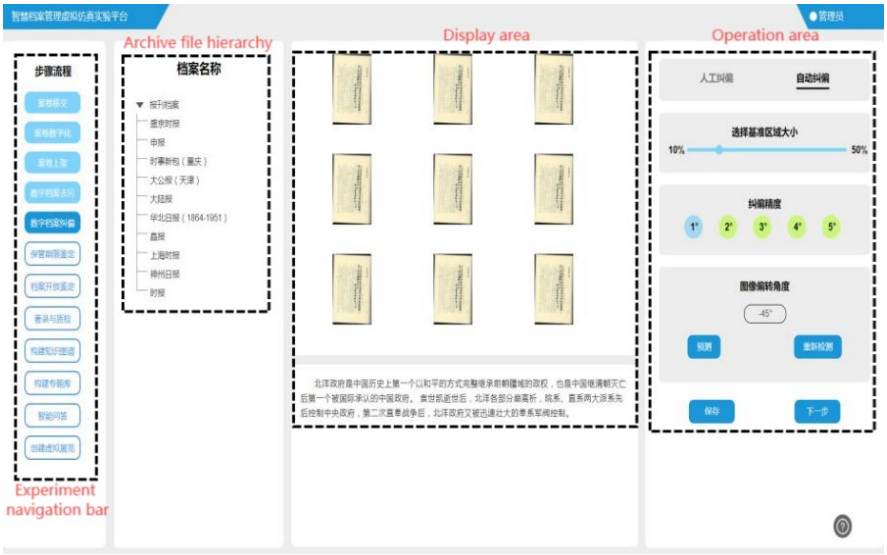
Intelligent Correction of Digital Archives: In manual correction, students select algorithms, adjust the 'detection accuracy adjustment' slider, preview the correction level, and save upon completion. In the automatic correction phase, students select correction algorithms and adjust the 'reference area size' slider to choose an appropriate reference area. The algorithm calculates the offset based on the selected reference area, and students select one of five accuracy options based on the offset size. Students can preview the correction results; if the reference area is inappropriate or the correction accuracy is too high, leading to poor results, they can click 're-detect' and proceed to the next step upon completion. This approach enhances correction accuracy using AI while combining AI and manual methods to ensure quality and efficiency.

Retention & Disposal: The experimental platform offers automatic classification and auxiliary retention& disposal prediction based on information fusion. An OCR recognition model is presented. The text parameters and sensitive words can be adjusted to modify the algorithm parameters in an interactive way to obtain different retention period suggestions. In addition, students can make final decision on how long an archival file should be retained or be disposed based on algorithmic results in an AI-human collaboration manner. This process allows students to participate in practical applications of retention period prediction and further optimize algorithms to improve prediction accuracy and efficiency.

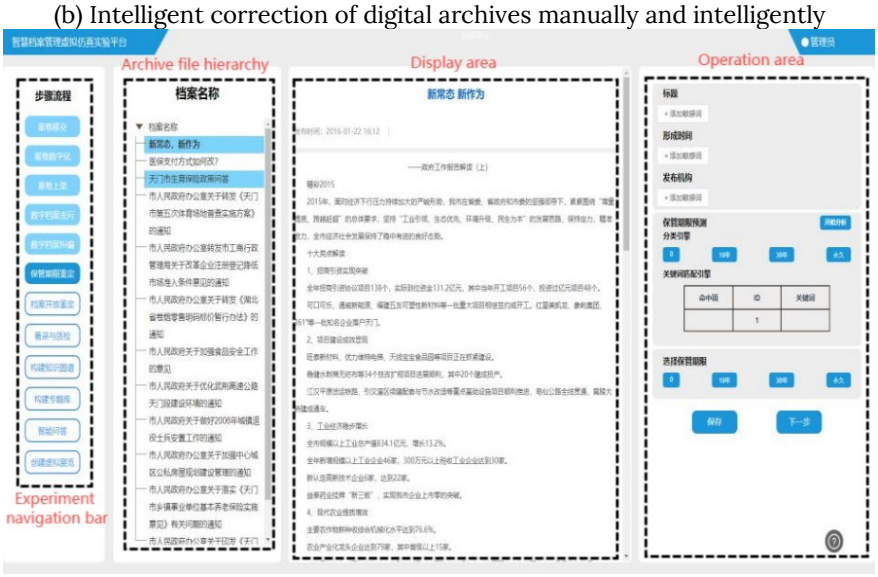
Metadata registration: On one hand, students manually register metadata such as titles, authors, institutions, and abstracts by checking archive file contents per-file one by one. As a comparison, an OCR based data recognition approach is presented where the metadata items are automatically recognizes. Students can choose to save manual or automatic registration results based on accuracy.



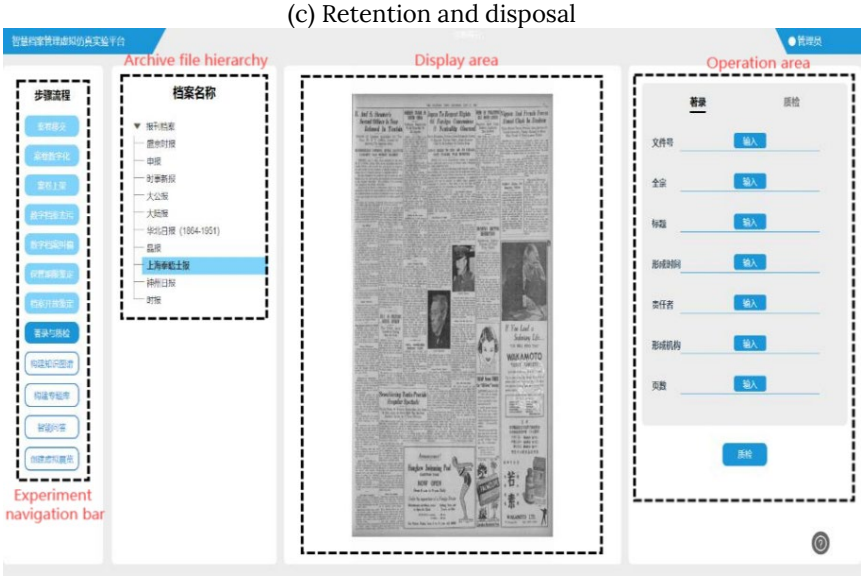
(a) Decontamination of digital archives manually and intelligently



(b) Intelligent correction of digital archives manually and intelligently



(c) Retention and disposal



(d) Metadata registration

Figure 4. Module 2-Interactive AI-models with adjustable parameters during archive management workflow

Module 3: AI-powered online archive exhibition automation

There are many approaches in which digital humanity applications can be developed from archives. However, with the assistance of AI technologies, a more intelligent development aiming at online exhibitions automation was proposed in this module. In addition, the experimental platform demonstrates multidimensional outcomes of AI-empowered archival development through constructing knowledge graphs, building thematic databases, developing a Q&A system from archive repository, and creating virtual exhibitions.

Knowledge graph building: the purpose of building a knowledge graph is to extract data and relate them from independent archive files for knowledge discovery etc. It is an important and intermediate step for developing digital humanity applications from archives. An ontology-based method is presented for students to build knowledge graph from archive files simply because we think it is quite a classic way to inform the students the basic structure of a knowledge graph other than large language model-based knowledge graph generation. By designating a topic beforehand, students construct their own ontology models by defining concepts, relationships, and properties. To complete the knowledge graph, specific data from archive files are extracted. SAMPP prefers to have students extract the data manually so they can understand how a knowledge graph is constructed step by step. The recognized archive data and metadata registered in previous steps are linked to the defined concepts and relationships for the completion of the themed knowledge graph. SAMPP also contains a built-in knowledge graph, and students' constructed graphs are compared with this built-in graph for similarity scoring.

Archive clustering: archive clustering is to group archive files with similar topics or contents. It is considered a downstream task of knowledge graph building. Based on specific keywords, combined with semantic similarity calculation, students can precisely locate semantically related nodes in the knowledge graph. Using these nodes, students can identify thematically and semantically similar topics and group the corresponding archival files. These files are then put into the thematic database. As a preparation, the clustered archive files can be put in the same area when setting up the online archive exhibition.

Q&A: on one hand, the Q&A task is a supplementary to archive clustering in case that some archive files are missing due to the natural imperfection of computer algorithms. On other hand, it is also a way to 'let the archive talk' and fill in the gap between the public and the mysterious archive files. The purpose of such an experimental task is to help students learn some fundamentals about data retrieval. The pre-built knowledge graph also plays a significant role in this task. With the semantic matching algorithms, students identify keywords from the textual questions and adjust parameters to locate candidate answer nodes in the knowledge graph. Once candidate answer nodes are found, students perform further aggregation analysis to synthesize and refine the information, ultimately generating responses in natural language.

Online 3D Exhibition: currently, online exhibition is a widely used way to show collections from museums, archives and library to the public(Kuo et al., 2013). That's why we chose to have online exhibition as the last but not the least task in SAMPP. As the foundations laid by the previous steps, archives files are grouped under a certain theme, which can also be the theme of our online virtual exhibition. A virtual exhibition template design component is given so the exhibition style and layout can be decided but also an automatic tour, real-time positioning, and exhibit zooming for interaction optimization can be determined by the student. Such a process provides a chance for students to integrate their creativity and appreciation of the beauty in showing the historical and cultural values from archive files.

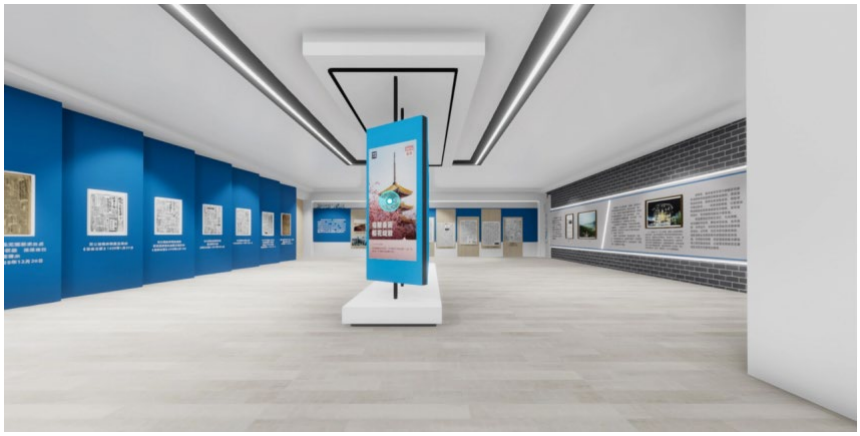
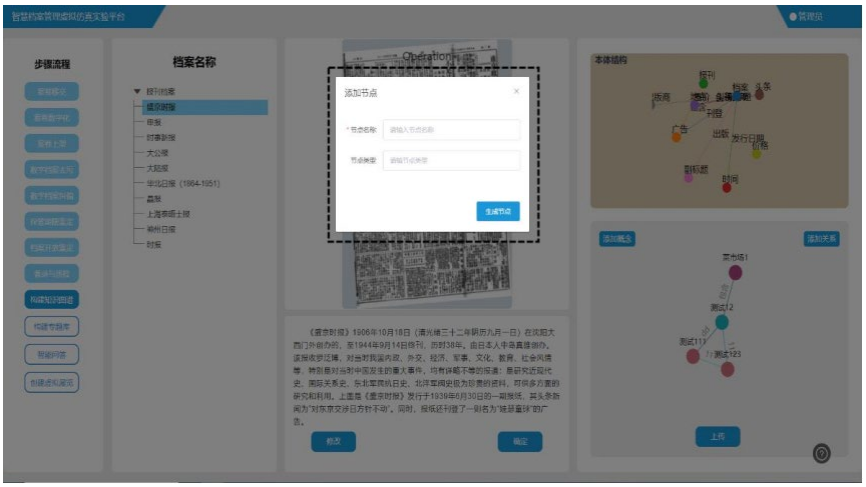
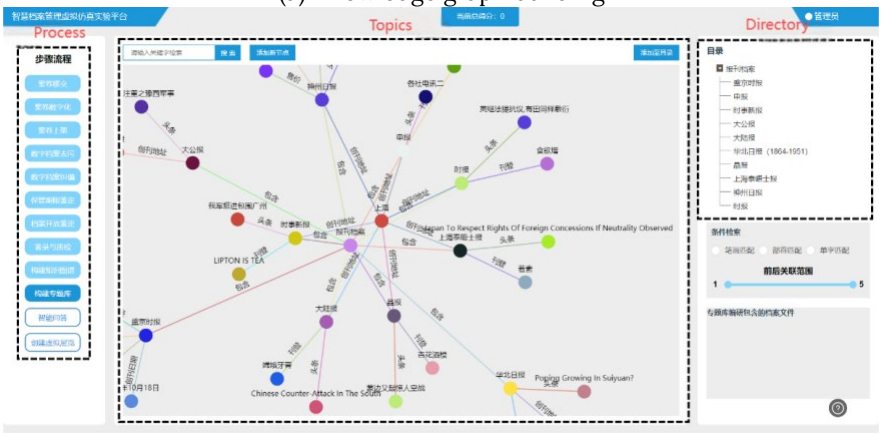


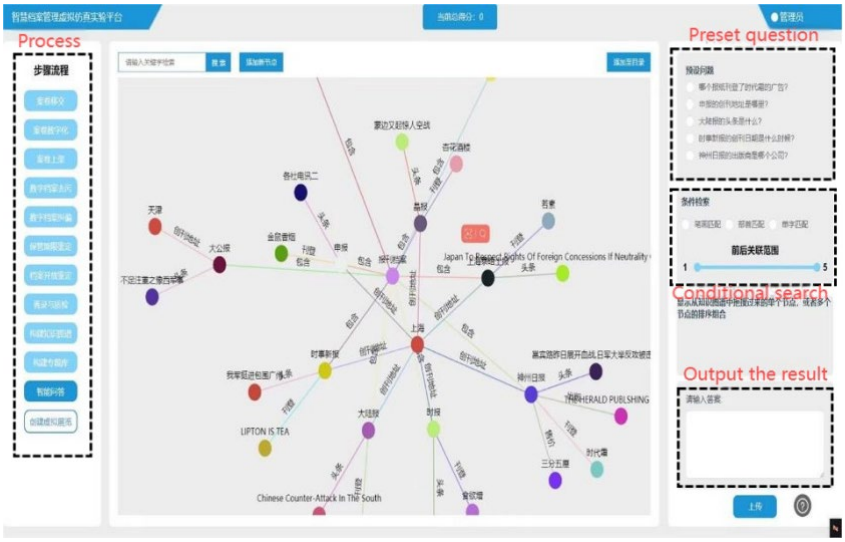
Figure 5. A Virtual exhibition hall of Wuhan urban memory



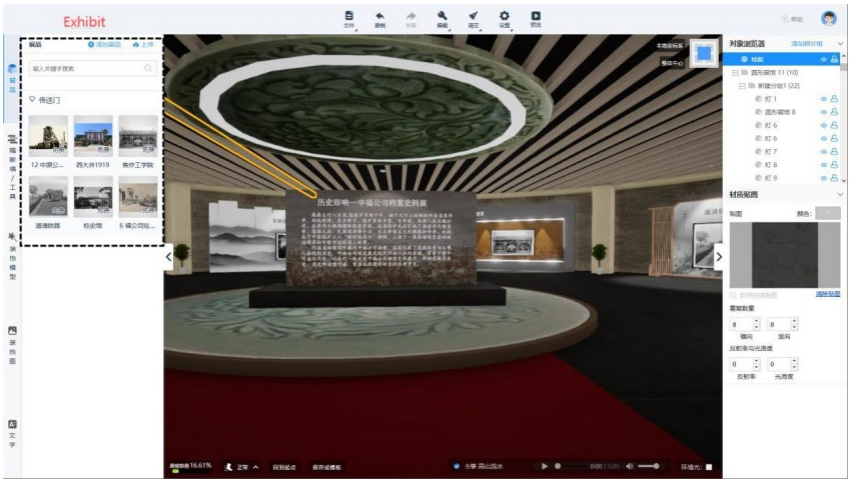
(a) Knowledge graph building



(b) Archive file clustering



(c) Developing a Q&A system



(d) Virtual exhibition design component

Figure 6. Module 3-Digital humanity application development from archive files

Student feedback

Sentiment analysis

By distributing questionnaires to undergraduates majoring in archival science at Wuhan University and conducting sentiment analysis on the responses, we compared the students' attitudes towards AI-learning before and after using SAMPP to explore the impact of the platform on students' attitudes and acceptance of technology.

Data collection: two separate questionnaires were designed for pre- and post- experiment survey, focusing on the understanding of AI technology and its application in the field of archives. The questionnaires were distributed to 55 undergraduates majoring in archival science at Wuhan University. The pre-experiment questionnaire, shown on the left column in Table.1, consists of three open questions: 1) The basic understanding of AI; 2) Their impressions of archive management work; and 3) Their confidence in learning and applying AI technology in the field of archival science. The post-experiment questionnaire, shown on the right column in Table.1, is also comprised three open questions: 1) Their understanding of integrating AI technology into the field of archives after completing the experiment; 2) The most impactful AI technology; and 3) Their

confidence in learning smart technology and applying it to archive management after completing the experiment.

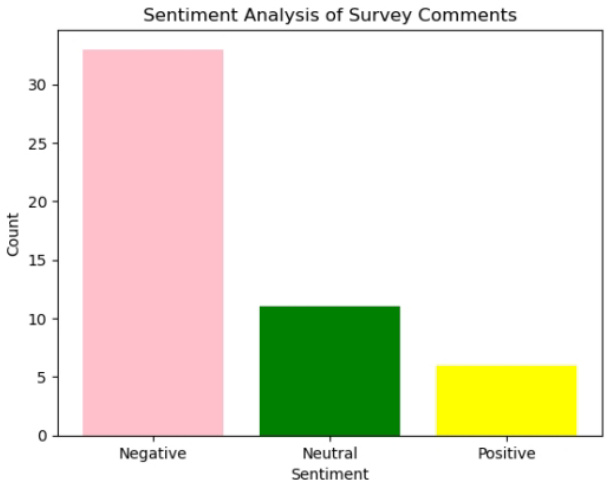
Name	Name
Gender	Gender
College Entrance Exam Major	College Entrance Exam Major
Question 1: What is your understanding of artificial intelligence before completing the experiment? Please describe in no less than 50 words.	Question 1: After completing the experiment, what is your understanding of the integration of artificial intelligence into the archival field? Please describe in no less than 50 words.
Question 2: Before completing this experiment, what was your impression of archive management? Please describe in no less than 50 words.	Question 2: Through this experiment, which type of smart technology or artificial intelligence technology did you experience the most? What is the reason?
Question 3: Before completing this experiment, do you have confidence in learning artificial intelligence technology? If not, what do you think is the biggest challenge? Please describe in no less than 50 words.	Question 3: After completing this experiment, do you have confidence in learning and applying artificial intelligence technology? Please explain the reason in no less than 50 words.

Table 1. The pre- and post- questionnaire

Data preprocessing: the collected questionnaires underwent a basic quality examination, excluding those not seriously answered or left unanswered, leaving 51 valid questionnaires. The responses were categorized by question and similar answers were merged. Given the unstructured nature and irregularity of the responses, they were segmented using the Jieba package for Chinese words tokenization, and data cleaning and preprocessing were performed using NLTK's stopwords, resulting in 306 usable sample data entries.

Sentiment Analysis: a sentiment dictionary is built for the sentiment analysis. For the third question in both pre- and post-experiment questionnaires, the students' confidence in learning and applying the technology was compared using sentiment analysis. Keywords expressing clear positive sentiments (‘*confidence*,’ ‘*like*,’ ‘*love*,’ ‘*determined*,’ ‘*certain*,’ ‘*confident*’) and negative sentiments (‘*no confidence*,’ ‘*dislike*,’ ‘*pessimistic*,’ ‘*hate*,’ ‘*worry*,’ ‘*confused*’) were identified. The domain sentiment dictionary in Python, SnowNLP was used for sentiment judgment training and testing. After satisfactory test results, the prepared dataset was imported for sentiment analysis, and visualized using matplotlib.pyplot, resulting in a sentiment bar chart.

Results: firstly, among the 51 pre-experiment questionnaires, 33 responses (66%) were categorized as negative emotions, 11 (22%) were neutral, and only 6 (12%) were identified as positive emotions, as it is shown in Figure.7.



Sentiment	Count	Percentage
Negative	33	66.0
Neutral	11	22.0
Positive	6	12.0

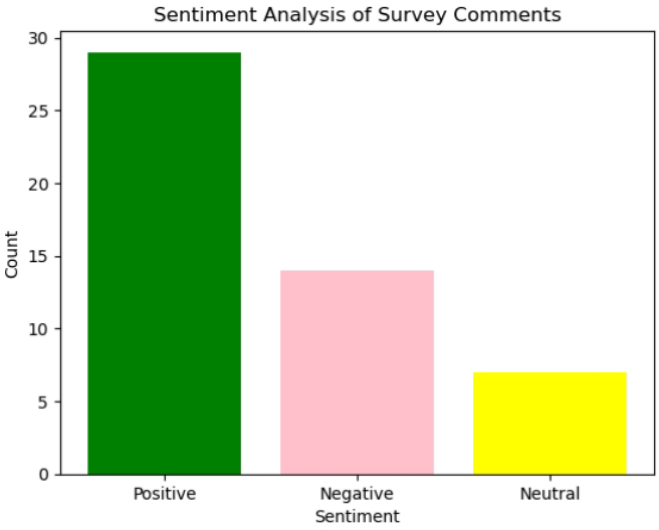
Figure 7. The sentiment analysis results from the pre-experiment questionnaire

Most students initially showed negative attitudes towards technology learning, considering AI techniques to be extremely difficult to master and beyond their capabilities. One student remarked, 'I think AI technology is very STEAM-based; it requires a lot of domain-specific knowledge and is very difficult to understand. As a liberal arts student, I may find it challenging.' Loads of similar responses indicated that undergraduates in archival science felt nervous and reluctant to accept technology due to their liberal arts background and preconceived notions about the complexity of technology.

However, after completing the experiment, the number of positive answers increased significantly. Many students felt that immersive, hands-on experience demonstrated the convenience of integrating AI technology into archive management practice and gained a deeper understanding of AI technology, as it is shown in Figure.8. Although 14 students still maintained a pessimistic attitude and 7 remained neutral, with mixed feelings of anticipation and anxiety over technology learning, the majority experienced a shift in attitude. Positive sentiments accounted for 58% of the responses (29 records). One student expressed her change in attitudes towards AI learning during the interview:

> *before completing the experiment, I thought that AI technology was extremely difficult to learn and master. But after the experiment, I deeply understood the significance of technology-enabled archival management. Not only did I build strong confidence in learning and applying technology, but I also released that it is an essential skill that we must master.*

This shows that instructive virtual learning can help students over stereotypes, develop new skills and a sense of closeness to technology, and become willing to learn and master and apply them to practical archive management work.



Sentiment	Count	Percentage
Positive	29	58.0
Negative	14	28.0
Neutral	7	14.0

Figure 8. The sentiment analysis results from the post-experiment questionnaire.

Deep interviews

Next, in order to further understand the students' perceptions after using the platform and its impact on their AI literacy, we conducted interviews. We randomly selected 22 students who had previously participated in the questionnaire survey and focused on two aspects: the application of AI technology during the experiment and the students' understanding of AI technology after the experiment.

Number	Perspective	Question
1	Application of AI technology in the experiment	In which stages of the experiment did you use AI technology?
2		What are the differences between using AI technology and not using it in these stages?
3	Understanding of AI technology	What role do you think AI technology plays in archive management after completing the experiment? Why?
4		Would you consider using AI technology in actual management work in the future?

Table 2. Questions for deep interviews

From the analysis of the interview results, it is evident that students could keenly perceive the use of AI technology during the process. A small number of students could even specify the technologies used in tasks such as decontamination, registration, and application development. Most students perceived differences in archive management with and without AI technology, with the consensus being that AI improves the efficiency and quality of archive management, especially in repetitive tasks where AI can complete them more accurately and quickly. One student responded:

> *in most stages, we need to complete tasks using both intelligent and manual methods. Comparing the two, I can clearly feel the convenience of intelligent technology-enabled management. Especially in the rectification stage, if I were to correct errors manually, the results would be poor. I might miss many issues and not know how to fix them. However, using technology, I found that if I set the accuracy high, it could quickly and comprehensively address the issues, far better than my manual corrections.*

Such an attitude was not isolated; 18 out of the 22 students believed that using AI technology is superior to purely manual methods, enabling quicker problem identification and resolution, and resulting in higher scores during the task evaluation phase. Regarding their understanding of AI, nearly 60% of the students viewed AI technology as an assistant in archive management, helping humans identify, judge, and respond more quickly. One student noted:

> *it's clear that the use of AI can help us complete tasks better and improve efficiency. It acts like a smart and responsive assistant, handling complex and repetitive work, making preliminary judgments, and providing suggestions. I believe that if used correctly, AI technology can significantly enhance the quality of our management work.*

Students who maintained a strongly positive attitude towards AI technology before its actual application believed in its widespread promotion to aid human decision-making. However, some students were cautious, arguing that in certain aspects, the use of technology might be counterproductive. They speculated that the excellent performance of AI in the experiments might be due to controlled settings and adjustments. In real-world scenarios, where archive management is more complex, the outcomes might not be as expected. One student said:

> *archives cover a wide range of content. Apart from regular and formatted document archives and specialized archives, most contain diverse societal content. AI technology struggles to possess such a comprehensive knowledge base to judge the authenticity of archives. In the identification stage of these archives, AI technology may not be reliable.*

Some students also expressed concerns about privacy leaks or security issues with extensive use of AI technology, advocating for a moderate and controlled application of technology. This indicates that students have a comprehensive understanding of AI technology in archive management, recognizing both its advantages and uncertainties, reflecting the platform's success in fostering AI literacy while also highlighting its design flaws:

- The knowledge base is highly predetermined, with many tasks preset. AI can leverage the prepared knowledge base to respond effectively, showcasing impressive results. However, in practical applications, the primary challenge is the incompleteness of the knowledge base, which students might not perceive during practice, leading to overly aggressive attitudes toward AI adoption.
- The experimental process is cumbersome, with numerous designed stages and tasks, making the experiment lengthy. Many students struggled to maintain focus in the later stages, showing impatience with manual problem-solving, potentially skewing their judgments and choices. Additionally, the experiment cannot be saved midway, requiring completion in one go, which can cause fatigue in the later stages.

Conclusion

In this paper, a virtual online learning platform, the Smart Archive Learning Platform (SAMPP), for students lacking a STEAM course learning experience is developed. The development of this platform follows a progressive 3-level strategy. The strategy fully considers the educational background of students majoring in archival science and puts the 3D immersive learning at the first level. The second level emphasizes mastering the AI techniques for daily archive management routine and the third level highlights building the creativity in developing digital humanity applications using AI techniques. SAMPP has three modules which has a 1 to 1 mapping to the 3-step strategy. A pre- and post- experiment questionnaire and deep interview are carried out to verify the usefulness of SAMPP in buildup students' confidence in AI-learning and developing their AI skill and show positive results. In future, we will be delicate in putting more AI models into each task to let students know the state-of-art progress of AI technology.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China (No. 72374161); 2023 Provincial Teaching Reform Research Project of Hubei Higher Education Institutions 'Development and Application of Virtual Simulation Training Platform of Intelligent Archives for Digital-Intelligent Interaction'; First-class Undergraduate Specialty Construction Project of Wuhan University 'Archival Science Driven by 'Integrated' National Security Strategy'.

About the authors

Lanxi Dong is an undergraduate student at Wuhan University's School of Information Management, focusing on artificial intelligence and archival education. She can be contacted at lanxidong@whu.edu.cn

Shuning **Tang** is a graduate student at the School of Information Management, Wuhan University. Her research interests include archival resource development and artificial intelligence. She can be contacted at 327201220@qq.com.

Yuan **Cheng** (PhD) is an Associate Professor at the School of Information Management, Wuhan University. Her research interests include archival resource mining, representation, and utilisation. She is the corresponding author of this article. She can be contacted at cynthia@whu.edu.cn

Ping Wang (PhD) is a Professor at the School of Information Management, Wuhan University. His research interests include digital technology and document digitisation, intelligent document development and services. He can be contacted at wangping@whu.edu.cn

References

Black, N. B., George, S., Eguchi, A., Dempsey, J. C., Langran, E., Fraga, L., Brunvand, S., & Howard, N. (2024). A Framework for Approaching AI Education in Educator Preparation Programs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23069–23077.

Bouloukakis, M., Partarakis, N., Drossis, I., Kalaitzakis, M., & Stephanidis, C. (2019). Virtual reality for smart city visualization and monitoring. *Mediterranean Cities and Island Communities: Smart, Sustainable, Inclusive and Resilient*, 1–18. https://doi.org/10.1007/978-3-319-99444-4_1

Hartmann, C., Orli-Idrissi, Y., Pflieger, L. C. J., & Bannert, M. (2023). Imagine & immerse yourself: Does visuospatial imagery moderate learning in virtual reality? *Computers & Education*, 207, 104909. https://doi.org/10.1016/j.compedu.2023.104909

Ilerbaig, J. (2019). Records, Information and Data: Exploring the Role of Record-Keeping in an Information Culture. *The American Archivist*, 82(1), 221–225. https://doi.org/10.17723/0360-9081-82.1.221

Kuo, P.-J., Lein, Y.-N., Su, W.-H., Wang, Y.-T., Huang, Y.-N., & Chu, W.-C. (2013). Mobilizing 3D virtual artifacts exhibition system of National Palace Museum. *Archiving Conference*, 10, 66–69. 10.1109/UMEDIA.2015.7297443

Lu, J., Yang, X., Zhao, W., & Lin, J. (2022). Effect analysis of a virtual simulation experimental platform in teaching pulpotomy. *BMC Medical Education*, 22(1), 760. https://doi.org/10.1186/s12909-022-03836-3

Marsick, V. (2001). Informal and incidental learning. *The New Update on Adult Learning Theory/Jossey-Bass*. https://doi.org/10.1002/ace.5

Pinski, M., & Benlian, A. (2024). AI literacy for users–A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans*, 100062. https://doi.org/10.1016/j.chbah.2024.100062

Rao, D., & Stupans, I. (2012). Exploring the potential of role play in higher education: Development of a typology and teacher guidelines. *Innovations in Education and Teaching International*, 49(4), 427–436. https://doi.org/10.1080/14703297.2012.728879

Rossau, I. G., Skovfoged, M. M., Czapla, J. J., Sokolov, M. K., & Rodil, K. (2019). Dovetailing: Safeguarding traditional craftsmanship using virtual, reality. *International Journal of Intangible Heritage*, 14(2), 93. 10.35638/IJIH. 2019..14.011

Standardization, I. O. for. (2016). International Organization for Standardization. *Information and Documentation-Records Management. Part 1: Concepts and Principles*.

Wang, Z., Liu, Y., Wang, L., & Fu, L. (2024). A Mars Exploration Control Virtual Simulation Experiment Platform for Engineering Practice in Control Engineering Education. *IEEE Transactions on Education*. 10.1109/TE.2024.3392332

Wenjing, C. (2025). Simulation application of virtual robots and artificial intelligence based on deep learning in enterprise financial systems. *Entertainment Computing*, 52, 100772. https://doi.org/10.1016/j.entcom.2024.100772

Zhang, H. (2024). Application of system virtual simulation platform based on spectral sensors in home physical education curriculum. *Optical and Quantum Electronics*, 56(4), 502. 10.1007/s11082-023-06155-z

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

'Just because we can, does it mean we should?' The integrated data infrastructure in Aotearoa, and its implications for Māori

Daphne Boey, Jennifer Campbell-Meier, Spencer Lilley, and Janet Toland

DOI: https://doi.org/10.47989/ir30iConf47302

Abstract

Introduction. This paper discusses the importance of data in Māori society and Māori Data Sovereignty in Aotearoa New Zealand.

Method. A descriptive case study is presented of the Integrated Data Infrastructure (IDI), a trove of interlinked and de-identified data about people that is the cornerstone for evidence-based policy decisions in Aotearoa New Zealand. The case study highlights existing issues Māori have with regards to the collection and use of their data.

Analysis. The IDI is examined in relation to its historical and cultural context, taking into account the ongoing impacts of colonisation with respect to the ownership and management of data. Discussed are the safety mechanisms of the IDI, secondary data use, and social & cultural licence.

Conclusion. The paper concludes by acknowledging that the right of Māori to exercise their self-determination of how data is collected, stored, and accessed remains aspirational and there is a need for Māori to have greater control over the data and who has access to it.

Introduction

To Māori, tangata whenua (indigenous people) of Aotearoa New Zealand (hereafter Aotearoa), Māori data is '*information or knowledge that is about or from Māori people*' (TMR, 2018). This can include personal data, or information about language, culture, resources or environments. Māori data encompasses not just a broader range of types than Western data, but also ways of recording them (Kukutai et al., 2020).

Māori data is taonga (treasure), and this distinction is key to understanding Māori Data Sovereignty (MDS) as '*distinct to other nation-based notions of data sovereignty*' (Ruckstuhl, 2022). As taonga, data falls under the sovereignty of Māori as stated in Article 2 of Te Tiriti ('*tino rangatiratanga… o ratou taonga katoa*'), (Waitangi Tribunal, 2024). Data is '*relational, not just to an individual, but to the collective…[and] current generations act as kaitiaki*' (Ruckstuhl, 2022). MDS is thus a way for Māori to fulfil their role of kaitiaki (stewards) for data.

This paper discusses the importance of data in Maori society and looks at the impacts of colonisation and differing worldviews on that. This is done by highlighting some ethical concerns about the integrated data infrastructure (IDI), heralded by many as being world-leading (Mitchell, 2022): a data repository of aggregated administrative data from multiple sources, that forms the basis for evidence-based policy decisions in New Zealand. To that end, historical context is provided, together with a discussion of key concepts related to Māori data, followed by analysis of the IDI as a case study.

Historical context

Archaeologists determined the first Polynesian (who became Māori) arrivals in Aotearoa between AD 800 and 1300 (Waitangi Tribunal, 2011). Māori had a sophisticated knowledge system before the British made landfall: evidence of organised society, horticulture, tools and trade can be found (Wilson, 2020). Unruly settlers in the 19th century led residents of Aotearoa to raise issue with the Crown, which culminated in the signing of Te Tiriti O Waitangi (The Treaty of Waitangi) (Keane, 2012). Importantly, two versions exist, and differing translations are a key point of contention (Orange, 2023): The First Article in the Treaty, written in English, gave the Queen '*all the rights and powers of sovereignty*', while the Queen was given '*kāwanatanga katoa*' in Te Tiriti. The latter is also what most rangatira (Māori chiefs) signed (Waitangi Tribunal, 2024). Kāwanatanga more accurately translates to governance (Te Aka Māori Dictionary, n.d.) and thus, Article 1 asserts the right for the Crown to govern the land while remaining subject to Māori sovereignty (tino rangatiratanga).

This lays the foundation for much of the interaction between the Pākehā (European) system and its indigenous population, which extends to data as well. Divergent understandings between the peoples continued: Māori valued reciprocity, and land transactions under customary law not seen as '*sale*' but '*a transfer of particular rights which remained subject to Māori rights to the land*' (Boast, 2015). This was not the case, as the Land Acts of 1800s showed how Pākehā purchasing land in manners '*not necessarily conducive to the well-being of the indigenous population*' (Boast & Black, 2010) whittled the approximately 80% Māori-held land in 1860 to 27% in 1910, with the entire South Island '*in settler hands*' (Ministry for Culture and Heritage, 2021).

Key concepts

This section discusses concepts and ideas explored in the case study.

Māori data sovereignty (MDS)

MDS is about '*respecting Māori data and the ability for Māori communities to exercise power over data usage and outputs*' (Greaves et al., 2023). It '*supports the realisation of Māori and iwi aspirations… [and] is a key mechanism for enabling self-determination and innovation*' (Hudson, 2016).

There are parallels between data and self-determination, starting with the emphasis on '*sovereignty*' in both instances. Kukutai & Cormack (2020) assert '*Māori Data Sovereignty reflects a longstanding aspiration for rangatiratanga*'. They continue: '*Indigenous Data Sovereignty is both a critical enabler of Indigenous self-determination…and a powerful tool for Indigenous resurgence and decolonisation*'.

Due to the history of colonisation, Māori have little trust in the colonial system as subjects of data and anxiety about racial discrimination and how, while '*seemingly neutral, data could be applied for…the detriment of Indigenous peoples*' (Davis, 2016). There is also concern about the relevance of advice, support or policy made without understanding of the impacts of colonisation (Greaves et al., 2023) – or worse, in their absence (Davis, 2016). This reflects what Walter (2016), as cited by West et al. (2020), describes as the 5Ds: '*disparity, deprivation, disadvantage, dysfunction and difference; all of which increase the potential for collective stigmatisation*'.

Further, the loss of sovereignty and marginalised position Māori hold results in difficulty to influence misguided narratives made about them from the data, but also '*limited ability to engage in, or influence policies*' (West et al., 2020). The embedded assumption that '*West is best*' that justified colonisation, assimilation and discrimination is also what excludes Māori from key areas of the data ecosystem. Kukutai et al. (2021) terms it '*hierarchies of evidence*', which positions indigenous knowledge as '*lacking intellectual rigour or validity*'.

Challenges are compounded when considering the divergence in worldviews between the two groups: for Māori, sovereignty '*is always in a relational sense*' (Ruckstuhl, 2022), and that '*notions of property, ownership and privacy (Western in origin) are foreign to a normative and social system that emphasises totality and interconnectedness*' (Kukutai & Cormack, 2020).

This stems from Māori cosmogony, where all things descend from primal parents Ranginui and Papatūānuku (Hikuroa, 2017). Whanaungatanga (kinship; reciprocity) and whakapapa (genealogy; relationships) are defining principles from a Te Ao Māori perspective: everything is connected and can be traced back to Ranginui and Papatūānuku (highlighting whakapapa); our interconnectedness means the need for whanaungatanga to uphold the mauri (life force) of all we encounter. Importantly, Māori cosmogony also means that inanimate objects are not '*impersonal thing[s]… [everything] is united through this web of common descent*' (Waitangi Tribunal, 2011).

Closely related to the idea of upholding mauri is kaitiakitanga, or guardianship that arises from whakapapa and whanaungatanga – '*the obligation of kinsfolk to nurture or care for their relations*'. It relates not just to the physical act of caring, but also the caring for of '*spiritual aspects… [such as] mauri*' (Waitangi Tribunal, 2011). Māori fulfil their role as kaitiaki in regard to data in Aotearoa.

Data colonialism

Ultimately, the language around '*data unlock[ing] opportunities*' (NZ Data Futures Partnership, 2016) frames data as a natural resource and '*implicitly centres extractive logics*' (Kukutai & Cormack, 2020). Privileging Western approaches to harnessing the power of data and knowledge exacerbates this, resulting in '*a continuation of the processes and underlying belief systems of extraction, exploitation, accumulation and dispossession that have been visited on Indigenous populations through historical colonialism*' (Kukutai & Cormack, 2020).

Ruckstuhl (2022) references the assumption of terra nullius, or land without a master, which underpinned much of colonial logic. She also parallels that to data today: datum nullius, '*a blank slate on which could be constructed the edifice of a distorting "colonial archive"*'.

While Māori are not a monolith, colonisation was significant and traumatic. Data practices that centre around extractivism is seen as

> *a 'replaying' of a familiar colonial experience, whereby 'resources' are seen to be open for exploitation and extraction of profit, with little regard for Indigenous knowledge systems and ways of conceptualising benefit (Kukutai & Cormack, 2020).*

Case study: integrated data infrastructure

Statistics New Zealand (StatsNZ) was formed in 1956. As the national statistics office, StatsNZ not only produces statistics about New Zealand society, but is also the *'lead agency for government-held data'* (StatsNZ, 2024).

StatsNZ manages the IDI, a trove of interlinked and de-identified data about people that sits as the cornerstone for evidence-based policy decisions in Aotearoa. Formally established in 2011, the IDI began '*as a technical exercise in data aggregation*' (Sporle et al., 2020) and is the culmination of the incremental linking of datasets held by StatsNZ and government agencies at various times (Greaves et al., 2023). The IDI harnesses the power of data to '*provide insight into our society and economy*' (StatsNZ, 2022a) and drive '*data-driven innovation*' (NZ Data Futures Partnership, 2016).

Researchers use the IDI to conduct cross-sector research that provides insights into our society and economy. A notable example is COVID-19 spread modelling during the pandemic. The IDI has wide reach – data sources include StatsNZ, government agencies, and even researchers who apply to '*integrate their datasets to the IDI*' (Greaves, et al., 2023).

Despite its potential, the IDI suffers from low public awareness (Gulliver et al., 2018) and was criticised as having been '*established without due democratic process*' (Greaves, et al., 2023), as its creation was justified under the existing statistics act 1975 and privacy act 1993, meaning no new legislation – or its accompanying parliamentary process – was required.

Analysis

Safety mechanisms of the IDI

To minimise harm that could result from the misrepresentation of Māori, and as an example of the Crown honouring Te Tiriti, StatsNZ has a working relationship with both Te Mana Raraunga (TMR) and the Data Iwi Leaders Group (DILG). The former's purpose is to advance MDS as described in its Charter (Te Mana Raraunga, 2018), while the DILG is '*a subcommittee of the Iwi Chairs Forum, a platform for sharing knowledge and information amongst tribal authorities in Aotearoa*' (Sporle et al., 2020). As a network of professionals with their own experience and expertise in data, research, policy and MDS, TMR is well-placed to undertake operational work with government agencies and DILG technical advisors; DILG's political mandate enabled them to engage directly with Government with regards to Māori data governance (Sporle et al., 2020).

This resulted in the Mana Ōrite Relationship Agreement, where parties can '*realise the potential of data to make a sustainable, positive difference to outcomes for [Māori]*' (StatsNZ, 2021). It also led to the creation of Ngā Tikanga Paihere (Sporle et al., 2020) to guide appropriate use of data in the IDI, with a focus on how '*Māori and other underrepresented populations are used for research purposes*' (Jones et al., 2022).

Ngā Tikanga Paihere, designed to provide cultural protection, sits alongside a broader safety framework: '*Five Safes*', referring to conditions ensuring safe access to data (StatsNZ, 2022b). The former draws from Māori tikanga (cultural lore; guidelines) to derive good data practices '*through cultural connection*' and assesses the '*cultural appropriateness of research for Māori… urges prioritis[ing] Māori participation and community relationships*' (StatsNZ, 2022b). Five Safes '*provides access to integrated data if "five safes" conditions are matched*': people, projects, settings, data, and output (StatsNZ, 2022b). StatsNZ vets researchers who apply to use the IDI, limiting data access through '*a secure virtual environment in approved data labs*', and reviewing research outputs before they are released (Greaves et al., 2023).

It is arguable if these are sufficient (Sporle et al., 2020). Ngā Tikanga Paihere remains a guideline, and the enforced Five Safes framework provides high-level protection against the leak and misuse of data, but much of the effort revolves around confidentiality and safety to protect the individual identities from whom data originated (StatsNZ, 2022b). '*Safe people*' refers to researchers who are technically competent; while '*Safe projects*' are based on if a project is done '*in public interest*'. It's also uncertain if safe use outside of StatsNZ can be ensured, as '*there is no standard procedure for ensuring the ethical use of IDI data*' due to the lack of '*a centralised statutory body responsible for oversight of research ethics in New Zealand*' (West et al., 2020). Further, researchers beyond government institutions are granted access to the IDI if they pass Five Safes, such as think tanks – which can exacerbate these concerns.

The inclusion of Ngā Tikanga Paihere alongside Five Safes is welcome and necessary but does beg critical analysis of whether it can be effectively utilised by researchers, especially in the context of Māori perspectives devalued in favour of Western frameworks because of colonisation, addressed above.

Secondary data use

The IDI's operation is entirely in secondary data use, and aggregated data was '*collected as part of routine transactions with state agencies, by busy staff who may be untrained, and are seldom users of the data they collect*' (Greaves et al., 2023). The variety of sources from which data is aggregated means different '*context…[and] ways… by different people… for different systems and purposes*', which bring to question how appropriate it is to aggregate them. Administrative data not originally intended for research/policy '*may not reflect robust research processes*' (Greaves et al., 2023). Even StatsNZ recommends contextual use of data: '*using the right data in the right context can substantially improve decision-making and… avoid generating potentially harmful outcomes*' (StatsNZ, 2018).

There is also a '*lack of clear, established consent processes for secondary data use*' (Sporle et al., 2020). Further, additional datasets can be added to the IDI (StatsNZ, 2024). These are '*data not collected as official statistics, for example independent survey or NGO/independent provider data*' (Radio NZ, 2018 as cited by Greaves et al., 2023). The literature is clear on how survey or data collection design can impact the data that is collected (see for example, Lietz, 2010). Including datasets beyond official government statistics or administrative data opens the IDI up to exponential impact from concerns outlined above.

For Māori, these issues are compounded – the IDI concentrates existing issues Māori have with regards to the collection and use of their data, for example in deficit framing, being overrepresented (e.g. justice records) (West et al., 2020), or underrepresented due to lack of institutional trust (e.g. census) (Davis, 2016), or conclusions made about Māori deemed to be objective even though it neglects crucial historical or social context (Greaves et al., 2023).

Social and cultural licence

In lieu of actual consent for secondary data use, this permission is instead, said to '*rely on social license*' (Greaves et al., 2023), which leverages on the trust New Zealanders have for Statistics NZ to act as stewards (Sporle et al., 2020).

Trust and the social licence with respect to data kaitiakitanga is however complex when one factors in '*Aotearoa's context of ongoing colonisation and contemporary disparities*' (Greaves et al., 2023). This is despite the commitment to co-governance, and on top of various frameworks and safeguards implemented with Māori consultation, because of general mistrust ensuing expectations of '*non-performativity*' (Kukutai & Cormack, 2020). As Greaves et al. (2023) note, '*the power… sits with the state, which minimises the autonomy and control that Māori can exercise.*' Social licence is also not a static concept and can be rescinded any time (Hudson, 2016).

TMR is critical about the use of social licence and clarifies this is due to the lack of distinction '*between individual and community acceptability of data use and sharing*' in addition to the fact that social license is informal. To them, social licence only seeks consent from individuals; a cultural licence should be extended to obtain permission *'based on the trust that iwi and Māori Treaty Partners have*'. This is because '*group acceptance through mandated structures is a more appropriate barometer of trust for data that can be aggregated to represent a group*' (Te Mana Raraunga, 2017).

Conclusion

The right of Māori to exercise their self-determination of how data is collected, stored, and accessed remains aspirational. Although users of the IDI are expected to adhere to Five Safes and Ngā Tikanga Paihere, this necessitates a high trust environment, which strengthens the argument for a need for Māori to have greater control over the data and who has access to it. Further, this enables Māori to fulfil kaitiaki responsibilities and ensure their perspectives and worldviews remain intact.

In Aotearoa, Te Tiriti paved the way for Crown-Māori relations that support '*the right to self-determination*' as declared in United Nations Declaration on the Rights of Indigenous Peoples (United Nations, 2007). With regards to data, this relationship has led to formal partnerships with indigenous advocacy groups, the development of a high-level data governance model and IDI-specific protections. Despite that, fears about data colonialism are warranted, advances towards bicultural governance have been critiqued for its '*non-performativity*' (Kukutai & Cormack, 2020).

This case study on the IDI is two-pronged – to highlight potential problematics of the platform and how these issues are exacerbated by historical context. This is especially so when considering the IDI's origins as the tool that enabled the National government's Social Investment, a transactional approach that sees its citizens as numbers on a page to manipulate (Baker & Cooper, 2018). Data, inherently non-neutral, can become weaponised in policy, and trickle-down effects could occur, such as changing the nature and approach of social service providers (e.g. NGOs), where services are built around funding, which is directed by government objectives – in this case, investment rather than protection (O'Brien, 2020). Additionally, questions about who needs intervention, the legitimacy of data practices, and the politics of data need to be answered (Baker & Cooper, 2018), but by whom? These are considerations that necessitate greater research into current operations of the IDI, to determine if the rights of New Zealanders, particularly Māori are impinged, and if the platform is ultimately boon or bane.

About the authors

Daphne Boey is an early-career scholar who has just graduated with a Masters in Business (Professional) at Victoria University of Wellington, New Zealand. Specialising in applied data analytics and sustainability, her focus areas intersect ethics, sustainability, data, government and policy. She is currently researching on Māori data sovereignty, and implications for bicultural environmental governance. A mid-career transition into academia, Daphne is in the process to undertake her doctorate in 2025.

Jennifer Campbell-Meier is Senior Lecturer in the School of Information Management, Te Herenga Waka - Victoria University of Wellington and former Programme Director for Information Studies. Jennifer published in the areas of information behaviour, digital inclusion, and professional development.

Spencer Lilley is an Associate Professor at the School of Information Management, Victoria University of Wellington, New Zealand. His research interests include indigenous information

behaviour, and the indigenisation of galleries, libraries, archives, museums, and records management institutions.

Janet Toland is an Adjunct Professor at the School of Information Management, Victoria University of Wellington, New Zealand. Her research interests are social and ethical aspects of the history of computing and sustainable information systems.

References

Baker, T., & Cooper, S. (2018). New Zealand's social investment experiment. *Critical Social Policy*, 38(2), 428–438. https://doi.org/10.1177/0261018317745610

Bishop, D. (2016). Indigenous peoples and the official statistics system in Aotearoa/New Zealand. In T. Kukutai, & J. Taylor (Eds.), *Indigenous data sovereignty: Toward an agenda* (pp. 291-306). Australian National University Press.

Boast, R. (2015, July 1). *Te tango whenua – Māori land alienation*. Retrieved from Te Ara Encyclopedia of New Zealand: https://teara.govt.nz/en/te-tango-whenua-maori-land-alienation/print

Boast, R. &. Black (2010). The Native Land Court and the Ten Owner Rule in Hawke's Bay, 1866-1873: An Analysis. *Land Law and Judicial Governance in the South Pacific*, Vol XII, 169-212. https://www.wgtn.ac.nz/law/research/publications/about-nzacl/publications/special-issues/hors-serie-volume-xii,-2011/Boast_Black.pdf

Davis, M. (2016). Data and the United Nations Declaration on the Rights of Indigenous Peoples. In T. Kukutai, & J. Taylor, *Indigenous data sovereignty: Toward an agenda* (pp. 1-24). Australian National University Press.

Greaves, L. M., Latimer, C. L., Muriwai, E., Moore, C., Li, E., Sporle, A., Clark, T. C., Milne, B. J. (2023). Māori and the Integrated Data Infrastructure: an assessment of the data system and suggestions to realise Māori data aspirations. [Te Māori me te Integrated Data Infrastructure: he aromatawai i te pūnaha raraunga me ngā marohitanga e poipoia ai ngā wawata raraunga Māori]. *Journal of the Royal Society of New Zealand*, 54(2), 190–206. https://doi.org/10.1080/03036758.2022.2154368

Gulliver, P., Jonas, M., McIntosh, T., Fanslow, J., & Waayer, D. (2018). Qualitative research: Surveys, social licence, and the integrated data infrastructure. *Aotearoa New Zealand Social Work*, 30(3), 57-71. https://search.informit.org/doi/10.3316/informit.099710734019215

Hikuroa, D. (2017). Mātauranga Māori - the ūkaipō of knowledge in New Zealand. *Journal of the Royal Society of New Zealand*, 47(1), 5-10. doi: https://doi.org/10.1080/03036758.2016.1252407

Hudson, M., Farrar, D., McLean, L. (2016). Tribal data sovereignty: Whakatōhea rights and interests. In Kukutai, T. & Taylor, J., *Indigenous data sovereignty: Toward an agenda* (pp. 157-178). Australian National University Press.

Jenkins, K. (2018). Can I see your social license please?. *Policy Quarterly*, 14(4). https://doi.org/10.26686/pq.v14i4.5146

Jones, C., McDowell, A., Galvin, V., & Adams, D. (2022). Building on Aotearoa New Zealand's Integrated Data Infrastructure. *Harvard Data Science Review*, 4(2). https://doi.org/10.1162/99608f92.d203ae45

Keane, B. (2012, June 20). He Whakaputanga – Declaration of Independence. Retrieved from Te Ara - the Encyclopedia of New Zealand: https://teara.govt.nz/en/he-whakaputanga-declaration-of-independence/print

Kukutai, T.; McIntosh, T.; Durie, M.; Boulton, A.; Foster, M.; Hutchings, J.; Mark-Shadbolt, M.; Barnes, M.; Moko-Mead, T.; Paine, S.; Pitama, S.; Ruru, J. (2021). *Te pūtahitanga: A Tiriti-led science-policy approach for Aotearoa New Zealand*. Ngā Pae o te Māramatanga.

Kukutai, T., Cormack, D. (2020). "Pushing the space": Data sovereignty and self-determination in Aotearoa NZ. In Walter, M., Kukutai, T., Russo Carroll, S., Rodriguez-Lonebear, D, *Indigenous Data Sovereignty and Policy* (pp. 21-35). Routledge.

Kukutai, T., Carroll, S. R., & Walter, M. (2020). Indigenous data sovereignty. In *International Work Group for Indigenous Affairs, The Indigenous World 2020* (pp. 654-662). Eks-Skolen Trykkeri, Copenhagen, Denmark.

Lietz, P. (2010). Research into Questionnaire Design: A Summary of the Literature. *International Journal of Market Research*, 52(2), 249-272. https://doi.org/10.2501/S147078530920120X

Mc Cartney, A. M., Anderson, J., Liggins, L., Hudson, M. L., Anderson, M. Z., TeAika, B., Geary, J., Cook-Deegan, R., Patel, H. R., & Phillippy, A. M. (2022). Balancing openness with Indigenous data sovereignty: An opportunity to leave no one behind in the journey to sequence all of life. *Proceedings of the National Academy of Sciences*, 119(4), https://doi.org/10.1073/pnas.211586011

Mitchell, C. (2022, July 31). The Government has built a data colossus - is it playing with fire? Retrieved from Stuff.co.nz: https://www.stuff.co.nz/national/300630922/the-government-has-built-a-data-colossus--is-it-playing-with-fire

Ministry for Culture and Heritage. (2021, 10 20). Pressure on Māori land. Retrieved from Manatū Taonga — Ministry for Culture and Heritage: https://nzhistory.govt.nz/war/taranaki-wars/pressure-on-maori-land

NZ Data Futures Partnership. (2016). *Data Futures Partnership*. Retrieved from NZ Data Trust: https://static1.squarespace.com/static/602c4087e88200661a7fb30b/t/6034207747128d7c43043a4f/1614028925133/2015+6+DFF+Data+Futures+Partnership.pdf

O'Brien, M. (2020). Social Investment in Aotearoa/New Zealand: Meaning and Implications. *Social Sciences*, 9(7), 1-14. https://doi.org/10.3390/socsci9070111

Orange, C. (2023, March 23). *Te Tiriti o Waitangi – the Treaty of Waitangi*. Retrieved from Te Ara - the Encyclopedia of New Zealand: https://teara.govt.nz/en/te-tiriti-o-waitangi-the-treaty-of-waitangi/print

Ruckstuhl, K. (2022). Data Is a Taonga: Aotearoa New Zealand, Māori Data Sovereignty, and Implications for Protection of Treasures. *NYU J. Intell. Prop. & Ent. L.*, *12*, 391.

Statistics NZ. (2024, June). *Data in the IDI*. Retrieved from Statistics NZ Tatauranga Aotearoa: https://www.stats.govt.nz/assets/Uploads/June-2024-Data-in-the-IDI.pdf

Statistics NZ. (2022a, August 23). *Integrated Data Infrastructure*. Retrieved from Statistics NZ Tatauranga Aotearoa: https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/

Statistics NZ. (2022b, August 23). *How we keep integrated data safe*. Retrieved from Statistics NZ Tatauranga Aotearoa: https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/

Statistics NZ. (2021, February 12). *Mana Ōrite Relationship Agreement*. Retrieved from Statistics NZ Tatauranga Aotearoa: https://stats.govt.nz/about-us/what-we-do/mana-orite-relationship-agreement/

Statistics NZ. (2020, November 23). *Ngā Tikanga Paihere*. Retrieved from Data.govt.nz: https://data.govt.nz/toolkit/data-ethics/nga-tikanga-paihere/

Statistics NZ. (2018, May). *Principles for the safe and effective use of data and analytics*. Retrieved from Statistics NZ Tatauranga Aotearoa: https://www.stats.govt.nz/assets/Uploads/Data-leadership-fact-sheets/Principles-safe-and-effective-data-and-analytics-May-2018.pdf

Sporle, A., Hudson, M., West, K. (2020). Indigenous data and policy in Aotearoa New Zealand. In Walter, M., Kukutai, T., Russo Carroll, S., Rodriguez-Lonebear, D, *Indigenous Data Sovereignty and Policy* (pp. 62-80). Routledge.

Te Aka Māori Dictionary. (n.d.). *Te Aka Māori Dictionary*. Retrieved from Te Aka Māori Dictionary: https://www.maoridictionary.co.nz/

Te Mana Raraunga. (2018). *Māori Data Sovereignty Network Charter*. Retrieved from Te Mana Raraunga: https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5913020d15cf7dde1df34482/1494417935052/Te+Mana+Raraunga+Charter+%28Final+%26+Approved%29.pdf

Te Mana Raraunga. (2017). *Te Mana Raraunga Statement on Social Licence*. Retrieved from Te Mana Raraunga: https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5924791cbf629a1367a65717/1495562526902/TMR+-+Statement+on+Social+and+Cultural+License.pdf

United Nations. (2007). *United Nations Declaration on the Rights of Indigenous Peoples*. Retrieved from United Nations Department of Economic and Social Affairs Indigenous Peoples: https://www.un.org/development/desa/indigenouspeoples/declaration-on-%20the-rights-of-indigenous-peoples.html

Waitangi Tribunal. (2024, February 2). *About the treaty*. Retrieved from Waitangi Tribunal: https://www.waitangitribunal.govt.nz/en/about/the-treaty/about-the-treaty

Waitangi Tribunal. (2011). Ko Aotearoa Tēnei: Te Taumata Tuatahi. A report into claims concerning New Zealand law and policy affecting Māori culture and identity. Wellington, New Zealand: Legislation Direct.

Waitangi Tribunal. (2014). *He Whakaputanga me te Tiriti* (The Declaration and the Treaty): The Report on Stage 1 of the Te Paparahi o Te Raki Inquiry. Wellington, New Zealand: Legislation Direct.

West, K., Hudson, M., & Kukutai, T. (2020). Data ethics and data governance from a Māori world view. In Indigenous research ethics: claiming research sovereignty beyond deficit and the colonial legacy (pp. 67-81). Emerald Publishing Limited.

Wilson, J. (2020, April 1). *History*. Retrieved from Te Ara - the Encyclopedia of New Zealand: https://teara.govt.nz/en/history/print

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

The impact of privacy authorization framing in mobile apps on privacy disclosure: a regulatory focus approach

Zhuo Sun, Dandan Zhang, Xiaoning Gao, and Yan Jin

DOI: https://doi.org/10.47989/ir30iConf47299

Abstract

Purpose. In the digital age, the tension between mobile app users obtaining personalized services and granting privacy permissions has become increasingly pronounced. The behavioral drivers and mechanisms influencing privacy disclosure in the context of privacy permissions require further investigation. By focusing on users' privacy disclosure intentions, this study aims to assist app platforms in designing privacy permission content that aligns with consumers' diverse privacy preferences and psychological needs, providing theoretical guidance and references for efficient platform operation.

Method. The study integrates the framing effect theory and regulatory focus theory, utilizing a 2 (gain frame vs. loss frame) × 2 (promotion focus vs. prevention focus) between-subjects design with 120 participants. It investigates the interactive effects of privacy authorization framing and regulatory focus on privacy disclosure intentions and examines the mediating roles of perceived value and anticipated regret in this relationship.

Results. The interaction between privacy authorization framing and regulatory focus has a significant impact on consumers' privacy disclosure intentions.

Conclusions. When privacy authorization framing aligns with an individual's regulatory focus, it enhances perceived value and reduces anticipated regret, thereby further reinforcing privacy disclosure intentions.

Introduction

Mobile apps play a critical role in facilitating interactions between consumers and businesses in the digital age. By obtaining user consent for privacy permissions and applying data mining techniques, app providers can deliver personalized recommendations and targeted marketing. However, concerns about privacy violations have raised significant apprehension among users (Acquisti et al., 2020). To address this, app providers must strike a balance between leveraging data and protecting privacy within a regulatory framework by developing appropriate privacy consent strategies. Users' decisions to grant privacy permissions are influenced by how information is presented—this is known as the framing effect (Bandara et al., 2020). The framing effect can be categorized into two types: gain framing, which highlights the benefits of taking a particular action, and loss framing, which emphasizes the negative consequences of not doing so (Levin et al., 1998). In the privacy context, businesses can leverage the framing effect to enhance how they collect user data, thereby encouraging users to disclose personal information (Shore, 2022). However, current research rarely examines privacy disclosure behavior from the perspective of privacy permissions, particularly regarding how the framing of privacy consent impacts user decisions. Privacy disclosure refers to the act of users sharing personal information, and its underlying motivations can be better understood through regulatory focus theory. According to this theory, users' privacy decisions are often influenced by either a promotion focus, which emphasizes the potential benefits of sharing information (e.g., personalized services or enhanced user experiences), or a prevention focus, which prioritizes protecting information to avoid risks (Hung et al., 2024). Yet, the interaction between an individual's regulatory focus and privacy framing, and how this dynamic affects privacy disclosure behavior, remains an underexplored area that warrants further investigation.

To address the aforementioned research gaps, this study introduces the framing effect into the domain of privacy permissions, analysing how the interaction between individuals' regulatory focus and information framing influences their intentions to disclose personal information. Additionally, the study explores the mediating roles of perceived value and anticipated regret in this process. The findings aim to provide insights for mobile app providers on balancing data utilization with privacy protection, enhancing user engagement on platforms, and improving user retention.

Hypothesis development

The effect of matching authorization frame and regulatory focus on privacy disclosure intentions

According to framing effects theory, information frames can be categorized into gain frames and loss frames. Gain framing encourages proactive decision-making, while loss framing emphasizes strategies to avoid losses (Kim, 2023). However, consumer preferences and choices vary based on psychological tendencies, with individual sensitivity to information framing being influenced by regulatory focus (Zhang et al., 2022). When the information frame aligns with an individual's regulatory focus, its persuasive effect is enhanced. Specifically, in the context of privacy permissions, a gain frame in privacy authorization appeals to promotion-focused consumers by presenting high-quality services, which fosters a sense of self-fulfilment and encourages them to adopt a more accepting approach to privacy consent notifications. Conversely, a loss frame in privacy authorization influences prevention-focused consumers by highlighting potential losses from refusing to grant permissions, thereby eliciting a risk-averse response, and leading to greater acceptance of privacy consent notifications. Based on this, the following hypotheses are proposed:

> **H1.** The interaction between authorization framing and regulatory focus significantly affects consumers' privacy disclosure intentions.

H1a. Promotion-focused consumers exhibit higher privacy disclosure intentions in response to a gain frame compared to prevention-focused consumers.

H1b. Prevention-focused consumers exhibit higher privacy disclosure intentions in response to a loss frame compared to promotion-focused consumers.

The mediating role of perceived value and anticipated regret

Perceived value refers to consumers' psychological assessment of a product's benefits versus its costs, influenced by individual needs, expectations, and psychological states (Porter, 1985; Yang et al 2023). The level of perceived value directly affects individuals' motivation and decision-making (Zhu et al., 2020). Hayashi and Sasaki (2022) noted that aligning information framing with individuals' regulatory focus characteristics can influence decision-making strategies and subsequently enhance perceived value (Hayashi & Sasaki, 2022; Peng et al., 2021). In the context of privacy disclosure, users' decisions to disclose personal information involve a trade-off between the anticipated benefits and potential risks (Yang et al 2020). Therefore, this study considers perceived value as a measure of the balance between the app service experience and privacy disclosure risks, and proposes the following hypothesis:

H2. The perceived value of consumer information processing is higher when the authorization framing aligns with the consumer's regulatory focus.

H2a. Promotion-focused consumers exhibit a higher perceived value in response to gain frames compared to prevention-focused consumers, leading to an increased willingness to disclose privacy.

H2b. Prevention-focused consumers demonstrate a higher perceived value in response to loss frames compared to promotion-focused consumers, resulting in a greater willingness to disclose privacy.

Anticipated regret is a negative emotion arising from the expectation of future adverse outcomes or missed opportunities (Loomes & Sugden, 1982). It drives individuals to choose options that align with their goals to minimize regret during the decision-making process (García & Curras-Perez, 2019). Research indicates that promotion-focused consumers aim to maximize benefits, while prevention-focused consumers strive to minimize negative outcomes, with both seeking to reduce anticipated regret (Lim & Hahn, 2020). In the context of privacy authorization, anticipated regret is influenced by either a gain or loss frame, affecting consumers' privacy disclosure intentions through the interaction between privacy authorization framing and regulatory focus. For instance, promotion-focused consumers, when presented with a gain frame, are more inclined to agree to privacy permissions to achieve positive outcomes and thus experience lower anticipated regret, which enhances their motivation to disclose personal information. Based on this, the study proposes the following hypothesis:

H3. Consumer information processing yields lower anticipated regret when authorization framing matches the consumer's regulatory focus.

H3a. Promotion-focused consumers experience lower anticipated regret in response to gain frames, which leads to a higher willingness to disclose privacy compared to prevention-focused consumers.

H3b. Prevention-focused consumers experience lower anticipated regret in response to loss frames, which leads to a higher willingness to disclose privacy compared to promotion-focused consumers.

Based on the above hypotheses, the theoretical model for this study is proposed, as illustrated in Figure 1.

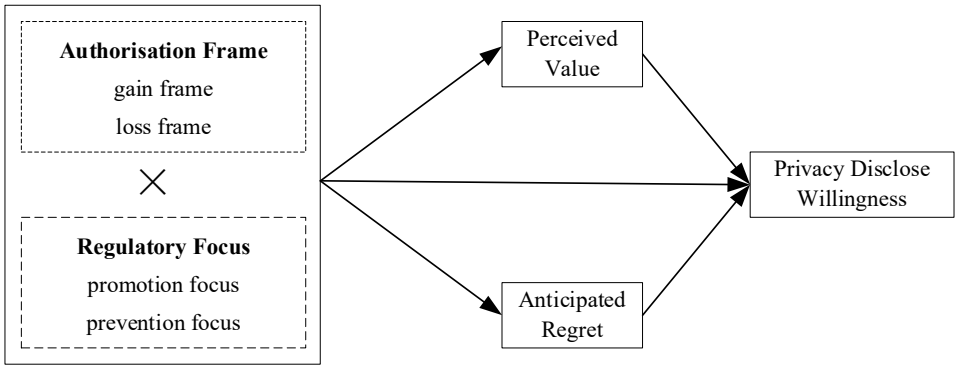


Figure 1. The theoretical model

Experimental design

Preliminary experiment 1

Preliminary experiment 1 aimed to identify the mobile app and type of privacy authorization to be used in the main experiment. First, social apps were selected based on their download numbers and ratings from the App Store and Google Play, as well as consumer preferences on social media platforms, focusing on apps with high user familiarity and clear core functionalities. Next, the privacy information requested by high-download and high-usage apps was analysed. Five types of privacy permissions commonly requested by social apps were identified: location, camera, photo library, microphone, and contacts. Finally, a survey was conducted with 102 university students (58 males, 44 females) who rated the perceived relevance of these five privacy permissions to the core functionalities of the apps using a 7-point Likert scale. The results showed that the contacts permission had the highest relevance, followed by photo library, microphone, camera, and location. Based on these results, the microphone permission was selected as the primary permission for mobile app requests in the main experiment to minimize the impact of the specific privacy permission on consumers' willingness to disclose personal information.

To eliminate the effects of participants' trust in existing apps and brand recognition on their willingness to disclose privacy (Bansal et al., 2016), this study designed a virtual social app named ‘Wetalk’. Two types of microphone authorization notifications were created: one emphasizing the benefits of using the microphone (gain frame) and the other highlighting the potential losses of not granting authorization (loss frame). As illustrated in Figures 2 and 3, the authorization notifications were presented as images in a uniform style at the bottom of the screen to simulate a real usage scenario and eliminate influences from location and style. Before the experiment began, participants were informed of the background setting: they were to assume they were interested in and had downloaded the ‘Wetalk’ social app, and upon preparing to use it, a privacy permission request would pop up on the screen.

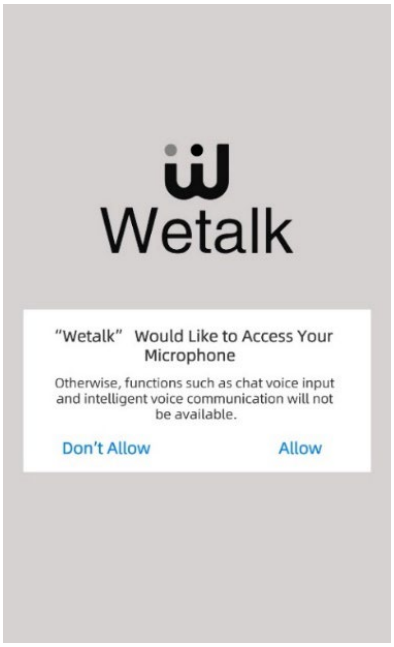


Figure 2. Gain frame



Figure 3. Loss frame

Preliminary experiment 2

Preliminary experiment 2 aimed to validate the effectiveness of the experimental stimuli in manipulating framing effects and regulatory focus. First, 60 students (33 males, 27 females) were randomly divided into two groups and shown the 'Wetalk' microphone permission authorization notifications framed either as a gain frame or a loss frame. After reading the authorization notifications, students rated the effectiveness of the framing. ANOVA analysis revealed significant differences between the two groups (P=0.000), with mean scores of 3.8 and 3.87, respectively, indicating that the framing manipulation was effective. Next, 112 students (54 males, 48 females) were randomly assigned to promotion-focused and prevention-focused experimental groups. Their regulatory focus was manipulated through questionnaire responses and short essay writing. ANOVA analysis showed that the scores for regulatory focus were significantly higher in both groups (P=0.000), with a mean score difference of 3.32 for the promotion-focused group and 3.25 for the prevention-focused group, demonstrating effective manipulation of regulatory focus. These manipulations provided effective experimental stimuli for the subsequent main experiment.

Formal experimental procedure

This study employed a 2 (gain frame vs. loss frame) × 2 (promotion focus vs. prevention focus) between-subjects design, involving a formal experiment with 120 students (63 males, 57 females) from a comprehensive university. Participants were randomly assigned to one of four experimental groups: gain frame - promotion focus, gain frame - prevention focus, loss frame - promotion focus, and loss frame - prevention focus. The mobile app 'Wetalk' and the validated microphone permission from preliminary experiment 1 were used as the privacy permission materials. The manipulation of framing effects and regulatory focus followed the validated methods from preliminary experiment 2. The experiment had two objectives: first, to examine the interactive effects of framing effects and regulatory focus on consumers' privacy disclosure intentions; and second, to test the mediating effects of perceived value and anticipated regret on consumers' privacy disclosure behavior.

In the formal experiment, a situational manipulation method was used, and participants were briefed on the experimental background before starting. First, the framing effects were manipulated by having participants view two types of authorization interfaces for the 'Wetalk'

microphone permission: gain frame and loss frame. Participants then evaluated the framing attributes of these interfaces. Next, the manipulation of regulatory focus was initiated through a '*short essay task*,' which included completing the Regulatory Focus Questionnaire and responding to items related to either promotion focus or prevention focus, as well as filling out a general regulatory focus measurement questionnaire. Finally, participants were asked to complete scales measuring their privacy disclosure intentions, perceived value, and anticipated regret based on their genuine responses. Additionally, information on their gender, age, educational background, and other individual characteristics was collected.

Measurement item design

To ensure the reliability and validity of the questionnaire measurements, this study drew on previous research and made appropriate modifications to the measures for privacy disclosure intentions, perceived value, and anticipated regret according to the specific research context. The experiment used a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree) to collect responses, with higher scores indicating a greater alignment with the survey items. For privacy disclosure intention, the measurement items included (Gabisch, & Milne, 2013; Wang et al., 2016): '*I am very likely to disclose my personal information to the mobile app*', '*I am willing to disclose my personal information to the mobile app*', and '*It is unlikely that I would disclose my personal information to the mobile app for the product's services*' (reverse-coded item). For perceived value, the measurement items included (Pham et al., 2023): '*disclosing my personal information to the mobile app would bring significant value*', '*the benefits of disclosing my personal information to the mobile app are highly attractive*', and '*disclosing my personal information to the mobile app is a good deal.*' For anticipated regret, the measurement items included (García, & Curras-Perez, 2019; Pham et al., 2023): '*if I refuse to disclose my personal information to the mobile app, I would regret missing out on the corresponding services*', '*once I make a decision, I don't look back*' *(reverse-coded item)*, and '*whenever I make a choice, I wonder what would happen if I had chosen differently.*' Finally, to control for individual characteristics that might influence privacy disclosure intentions, variables such as gender, age, and educational background were included as control variables.

Results

Main effect test

ANOVA was conducted to test Hypothesis H1. The results indicate that the interaction between authorization framing, and regulatory focus significantly affects privacy disclosure intentions. When the app requests user authorization under the gain frame, promotion-focused consumers (M=5.589, SD=0.615, F (1,118) = 234.342) and prevention-focused consumers (M=2.794, SD=0.653, F (1,118) = 234.342) show significant differences in their privacy disclosure intentions (P=0.000). This suggests that promotion-focused consumers exhibit higher privacy disclosure intentions compared to prevention-focused consumers within the gain frame. When the app requests user authorization under the loss frame, prevention-focused consumers (M=5.206, SD=0.653, F (1,118) = 267.522) and promotion-focused consumers (M=3.183, SD=0.701, F (1,118) = 267.522) also show significant differences in their privacy disclosure intentions (P=0.000). This indicates that prevention-focused consumers have higher privacy disclosure intentions compared to promotion-focused consumers within the loss frame. The impact of the alignment between information framing and regulatory focus on privacy disclosure intentions is illustrated in Figure 4, supporting Hypothesis H1.

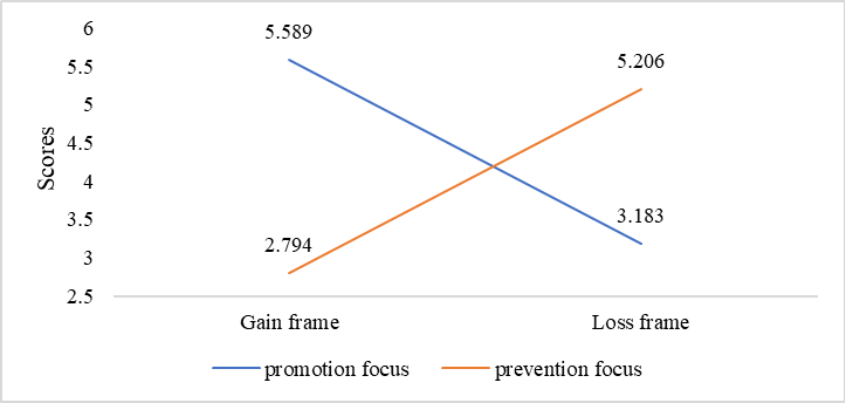


Figure 4. Effect of matching information framing and regulatory focus on willingness to disclose privacy

Mediation effect test

This study employed PROCESS to test the mediation effects. Bootstrap analysis results showed that for promotion-focused consumers, when the app requests authorization information under the gain frame, the confidence interval (BootLLCI = 0.5086, BootULCI = 0.9552) at a 95% confidence level does not include 0. This indicates that perceived value mediates the interaction between framing effects and privacy disclosure intentions. Similarly, for prevention-focused consumers, when the app requests authorization information under the loss frame, perceived value also serves as a mediator (BootLLCI = 0.1339, BootULCI = 0.4270). Additionally, ANOVA further confirmed that under the gain frame, promotion-focused participants (M=5.16, SD=0.612, F (1,118) = 26.189) reported higher perceived value compared to prevention-focused participants (M=4.55, SD=0.680, F (1,118) = 26.189) (P=0.000). Under the loss frame, prevention-focused participants (M=4.96, SD=0.552, F (1,118) = 125.139) reported higher perceived value compared to promotion-focused participants (M=3.53, SD=0.689, F (1,118) = 158.139) (P=0.001). The mean scores for perceived value under both gain and loss frames are illustrated in Figure 5, confirming Hypothesis H2.

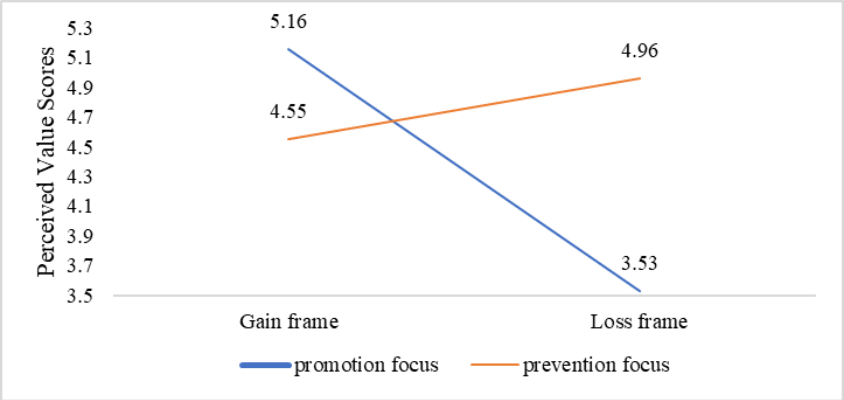


Figure 5. Perceived value scores

Next, the study analysed the mediating effect of anticipated regret within the gain and loss frames. Bootstrap analysis results indicated that for promotion-focused consumers, when the app requests authorisation information under the gain frame, the confidence interval (BootLLCI = 0.720, BootULCI = 0.4193) at a 95% confidence level does not include 0. This suggests that anticipated regret mediates the interaction between framing effects and privacy disclosure intentions. Similarly, for prevention-focused consumers, anticipated regret also serves as a mediator when the app requests authorization information under the loss frame (BootLLCI = 0.0621, BootULCI = 0.3045). Furthermore, ANOVA results showed that under the gain frame, promotion-focused participants (M=3.00, SD=0.559, F (1,118) = 235.104) reported lower anticipated

regret compared to prevention-focused participants (M=5.21, SD=0.968, F (1,118) = 235.104) (P=0.000). Under the loss frame, prevention-focused participants (M=3.17, SD=0.475, F (1,118) = 561.526) reported lower anticipated regret compared to promotion-focused participants (M=5.60, SD=0.635, F (1,118) = 561.526) (P=0.000). The mean scores for anticipated regret under both gain and loss frames are illustrated in Figure 6, confirming Hypothesis H3.

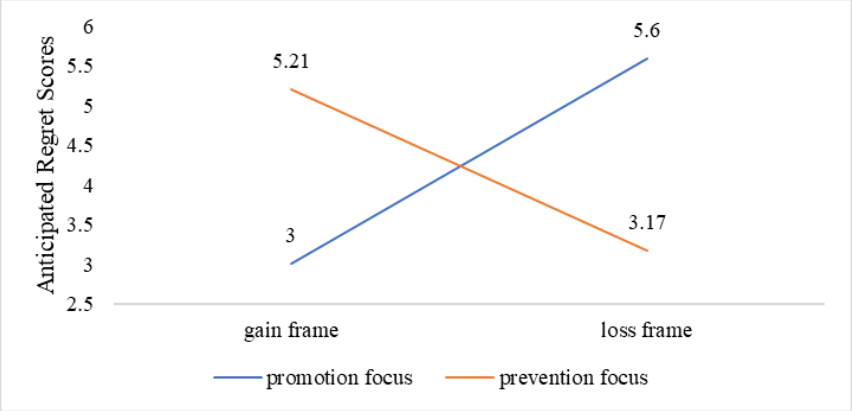


Figure 6. Anticipated regret scores

Conclusion

This study incorporates the framing effect into the field of privacy authorization to explore how the interaction between privacy authorization framing and consumers' regulatory focus influences privacy disclosure behavior. It also examines the mediating roles of perceived value and anticipated regret. The findings reveal that the interaction between privacy authorization framing, and regulatory focus significantly affects consumers' privacy disclosure intentions. When privacy authorization framing aligns with an individual's regulatory focus, it results in higher perceived value and lower anticipated regret, thereby strengthening privacy disclosure intentions. By offering a fresh perspective on how different descriptions of privacy authorization information impact user privacy disclosure intentions, this research expands the study of framing effects in mobile app privacy contexts.

This study also holds significant managerial implications. First, enterprises should refine the management of privacy authorization design by integrating user behavior and psychological analyses to optimize information presentation. This approach can enhance users' perceived value, reduce anticipated regret, and strengthen trust in privacy disclosure. Meanwhile, regulators could consider establishing dynamic oversight mechanisms to regularly assess and review privacy authorization practices, ensuring alignment with data protection standards while promoting innovation in privacy protection technologies. Furthermore, diverse feedback channels should be developed to encourage user participation in refining privacy frameworks, ensuring that privacy measures remain up-to-date and responsive to user needs. These measures will help safeguard user privacy, promote transparency, security, and user-friendliness in the mobile app industry, and support the healthy growth of the digital economy.

Acknowledgements

The research was supported by National Social Science Fund of China (23CTQ025), China Postdoctoral Science Foundation (2023M743213), Humanities and Social Sciences Research Project of Henan Province Colleges and Universities(2024-ZZJH-283) and Henan Key R&D and Promotion Project (Soft Science) (242400411171).

About the authors

Zhuo Sun is Associate Professor in the School of Information Management, ZhengZhou University, ZhengZhou, China. He received his Ph.D. from University of Chinese Academy of Sciences, and his research interests are privacy decision-making and social commerce. He can be contacted at sunzhuo@zzu.edu.cn

Dandan Zhang is a graduate student in the School of Information Management, ZhengZhou University, ZhengZhou, China. She is planning to apply for a PhD in the future and her research interests are users' information behavior. She can be contacted at dandanzhang0418@163.com

Xiaoning Gao is Associate Professor in the School of Information Management, ZhengZhou University, ZhengZhou, China. She received her Ph.D. from Xi'An JiaoTong University, and her research interests are data sharing decisions and privacy risk governance. She can be contacted at gxn2018@zzu.edu.cn

Yan Jin is Professor in the School of Information Management, ZhengZhou University, ZhengZhou, China. She received her Ph.D. from WuHan University, and her research interests are users' information behavior, information governance and health informatics. She can be contacted at jin_yan@zzu.edu.cn

References

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2020). Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. Journal of Consumer Psychology, 30(4), 736-758.

Bandara, R., Fernando, M., & Akter, S. (2020). Addressing privacy predicaments in the digital marketplace: A power - relations perspective. International Journal of Consumer Studies, 44(5), 423-434.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. Organizational behavior and human decision processes, 76(2), 149-188.

Shore, A. (2022). Talking about facial recognition technology: How framing and context influence privacy concerns and support for prohibitive policy. Telematics and Informatics, 70, 101815.

Hung, S. Y., Tsai, J. C. A., Chen, K., Chen, C., & Yeh, T. T. (2024). Tacit knowledge sharing in information systems development projects: social interdependence and regulatory focus perspectives. Information Technology & People, 37(4), 1449-1477.

Kim, J. (2023). Gain-loss framing and social distancing: temporal framing's role as an emotion intensifier. Health Communication, 38(11), 2326-2335.

Zhang, Z., Gong, M., & Jia, M. (2022). How and when does top management team regulatory focus influence firm environmental misconduct? human relations, 75(7), 1298-1326.

Porter, M. E. (1985) Competitive advantage: creating and sustaining superior performance. Competitive advantage creating sustaining superior performance, 25, 82-84.

Yang, J., Ren, J., Zhao, M., Chen, H., & Wang, H. (2023). The impact of value co-creation on consumer citizenship behavior: Based on consumer perspective. Frontiers in Psychology, 13, 1110011.

Zhu, P., Wang, Z., Li, X., Liu, Y. H., & Zhu, X. (2020). Understanding promotion framing effect on purchase intention of elderly mobile app consumers. Electronic Commerce Research and Applications, 44, 101010.

Hayashi, Y., & Sasaki, H. (2022). Effect of leaders' regulatory - fit messages on followers' motivation. Journal of Applied Social Psychology, 52(7), 496-510.

Peng, J., Cao, F., Zhang, Y., Cao, Y., Zhang, Y., Zhu, X., & Miao, D. (2021). Reflections on motivation: How regulatory focus influences self-framing and risky decision making. Current Psychology, 40, 2927-2937.

Yang, Q., Gong, X., Zhang, K. Z., Liu, H., & Lee, M. K. (2020). Self-disclosure in mobile payment applications: Common and differential effects of personal and proxy control enhancing mechanisms. International Journal of Information Management, 52, 102065.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. The economic journal, 92(368), 805-824.

García, I. S., & Curras-Perez, R. (2019). Is satisfaction a necessary and sufficient condition to avoid switching? The moderating role of service type. European Journal of Management and Business Economics, 29(1), 54-83.

Lim, J., & Hahn, M. (2020). Regulatory focus and decision rules: Are prevention-focused consumers regret minimizers? Journal of Business Research, 120, 343-350.

Bansal, G., Zahedi, F. M., & Gefen, D. (2016). Do context and personality matter? Trust and privacy concerns in disclosing private information online. Information & Management, 53(1), 1-21.

Wang, T., Duong, T. D., & Chen, C. C. (2016). Intention to disclose personal information via mobile applications: A privacy calculus perspective. International journal of information management, 36(4), 531-542.

Gabisch, J. A., & Milne, G. R. (2013). Self - disclosure on the web: Rewards, safety cues, and the moderating role of regulatory focus. Journal of Research in Interactive Marketing, 7(2), 140-158.

Pham, T., Mathmann, F., Jin, H. S., & Higgins, E. T. (2023). How regulatory focus – mode fit impacts variety - seeking. Journal of Consumer Psychology, 33(1), 77-96.

García, I. S., & Curras-Perez, R. (2019). Is satisfaction a necessary and sufficient condition to avoid switching? The moderating role of service type. European Journal of Management and Business Economics, 29(1), 54-83.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Does artificial intelligence harm labour? Investigating the limitations of incident trackers as evidence for policymaking

Theodore Dreyfus Ledford

DOI: https://doi.org/10.47989/ir30iConf47296

Abstract

Introduction. From the point of view of public policy, artificial intelligence (AI) is an emerging technology with as-yet-unknown risks. AI incident trackers collect harms and risks to inform policymaking. We investigate how labour is represented in two popular AI incident trackers. Our goal is to understand how well the knowledge organization of these incident trackers reveals labour-related risks for AI in the workplace, with a focus on how AI is impacting and expected to impact workers within the United States.

Data and Analysis. We search for and analyse labour-related incidents in two AI incident trackers, the Organization for Economic Cooperation and Development's AI incidents monitor (OECD AIM) and the AI incident database (AIID) from the responsible AI collaborative.

Results. The OECD AIM database categorised workers as stakeholders for 600 incidents with 6,744 associated news reports. From the AIID, we constructed a set of 57 labour-related incidents.

Discussion and Conclusions. The AI incident trackers do not facilitate ready retrieval of labour-related incidents: they used limited existing labour-related terminology. AI incident trackers' reliance on news reports risks overrepresenting some sectors and depends on the news reports' framing of the evidence.

Introduction

To make regulatory decisions, policymakers need to define problems and priorities in a process called agenda setting (Baumgartner, 2015). Issues raised in the agenda setting phase inform the evidence to be collected from empirical research (Head, 2010; Pawson et al., 2011) to shape the space of policies envisioned (MacKillop & Downe, 2023; Schiff, 2024).

Artificial intelligence (AI), while long studied in the academic sphere, is, from the point of view of public policy, an emerging technology with as-yet-unknown risks. In the public consciousness, two vivid risks people envision from AI are existential threats to humanity (Cameron, 1984) and the risk of being replaced by machines (Autor, 2015; 2022). In the past few years, multiple groups have introduced AI incident trackers (Abercrombie et al., 2024; Hutiri et al., 2024; McGregor, 2021; OECD, 2024) and taxonomies (Arda, 2024; Cattell et al., 2024; Critch & Russell, 2023; Shelby et al., 2022; Weidinger, 2022) to analyse the potential harms and risks of AI.

In this paper, we focus on labour relations with AI (Arntz et al., 2016), sometimes called the 'future of work' (Laker, 2023). Specifically, we investigate how two popular AI incident trackers represent labour-related risks for AI in the workplace.

Background

AI and labour

Policymakers seek to mitigate the effects of emerging technology on labour displacement (Lane & Saint-Martin, 2021), deskilling and economic trade competition (OECD, 2022b). In 2020, US policymakers defined an AI as a '*machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments*' (ScienceIsUS, 2024). As early as the 20th century, governments and research experts forecasted the effects of automation on the workforce (Acemoglu and Restreppo, 2019). More recently, Frey & Osborne (2017) measured which occupations are at the most risk of automation.

Incident tracking

We identified a proliferation of AI incident trackers (e.g., Hutiri et al., 2024; McGregor, 2021; Rodrigues et al., 2023; Shrishak, 2023). These AI incident trackers are informed by earlier incident reporting strategies to address system failures and risks in aviation (NASA Aviation Safety Reporting System, n.d.; Reynard, 1986), healthcare (Kohn et al., 2000; Macrae, 2016), software development (Booth et al., 2013) and cybersecurity (van der Kleij et al., 2022). However, current AI trackers '*rely heavily on news coverage of AI incidents*' (Turri & Dzombak, 2023). There is no standard structure for incident tracking. Incidents begin as records of an event that are deemed worth reporting. Incident tracking depends on a predefined documentation procedure indicating what information may indicate failure or risk within a system and is '*worth*' knowing (Turri & Dzombak, 2023). Recording an event as an incident transforms it into labelled data. Experts standardise the vocabulary that guides how incident trackers can assess the implications of a technology or policy (Hoffmann & Frase, 2023; OECD, 2022a).

Aims

To determine how well AI incident trackers inform policymaking, we study how well the knowledge organization of different incident trackers reveals labour-related risks for AI in the workplace. We search for and analyse labour-related incidents in two AI incident trackers, the AI incidents monitor and the AI incident database, with a focus on how AI is impacting and is expected to impact workers within the United States.

Data collection and analysis

AI incident tracker 1: OECD AI incidents monitor (AIM)

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organization of 38 mostly high-income, industrialised countries. The OECD AI incidents monitor, (AIM) launched publicly in 2023. OECD AIM was developed as part of their efforts on AI governance to '*establish a knowledge foundation… and… terminology*' for an interactive evidence base that could help policymakers to define the scope of AI (OECD, 2023). The incident tracker's backend process is supported by the event registry (http://eventregistry.org) digital service, a third-party commercial entity. Event registry monitors news reports worldwide, drawing on an expert-created category system and machine learning algorithms to group news into incidents and automatically classify harms. For OECD AIM, each incident is automatically assigned a summary and headline from the primary news report (the news report from the source with the highest Alexa traffic rank).

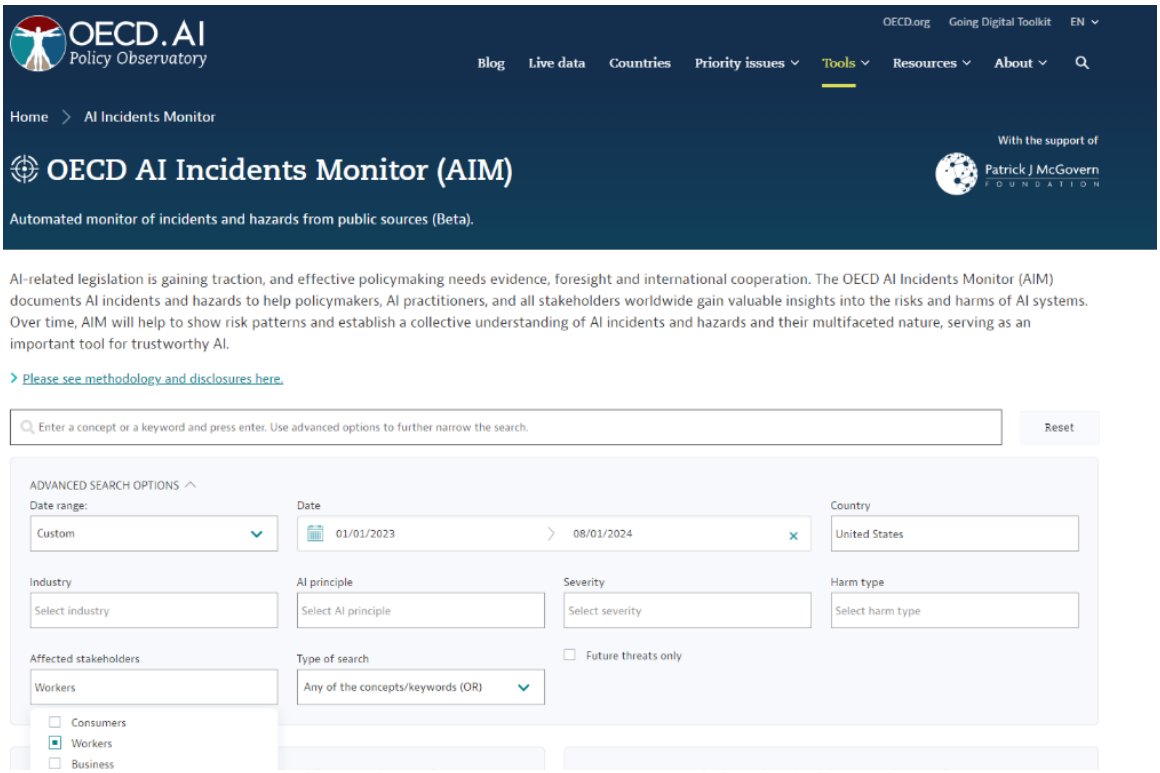


Figure 1. The OECD's AIM database query portal, showing some of our query settings

To collect data, we downloaded query results (incidents, summaries, headlines) from the OECD's AIM database (Figure 1) on August 1, 2024. We had to depend on the online portal because the most important data for our project, '*affected stakeholders*', was only available through this channel. (The batch data download option only provided the incident ID, title, summary, date of creation, concepts, and companies.) We always set '*affected stakeholders*' to workers, '*country*' to United States and '*date*' from January 1, 2023, through August 1, 2024.

Our queries retrieved 600 incidents with 6,744 associated news reports that the OECD AIM database categorised as having workers as '*affected stakeholders*'. We searched separately for each '*industry*' in the OECD AIM industry taxonomy and calculated the percentage of incidents with Workers as stakeholders. We searched with '*future threat only*' both unchecked and checked and calculated the percentage of future threats. We grouped industrial sectors by comparing the percentage of incidents that involved future threats, the percentage of incidents that involved

workers as stakeholders, and to what extent these intersected (e.g., workers as stakeholders AND future threats; neither; or just one or the other).

AI incident tracker 2: AI incident database (AIID)

The AI Incident Database (AIID) (https://incidentdatabase.ai) was launched in 2020 with support from the Partnership on AI (https://partnershiponai.org/) (McGregor, 2021). Drawing from crowdsourced submissions, the AIID depends on a taxonomy and annotation guidance developed by the Georgetown University Center for Security and Emerging Technology (Hoffmann et al., 2023; Responsible AI Collaborative, 2022). Each news report is assigned to a single incident, but each incident may collect multiple related news reports.

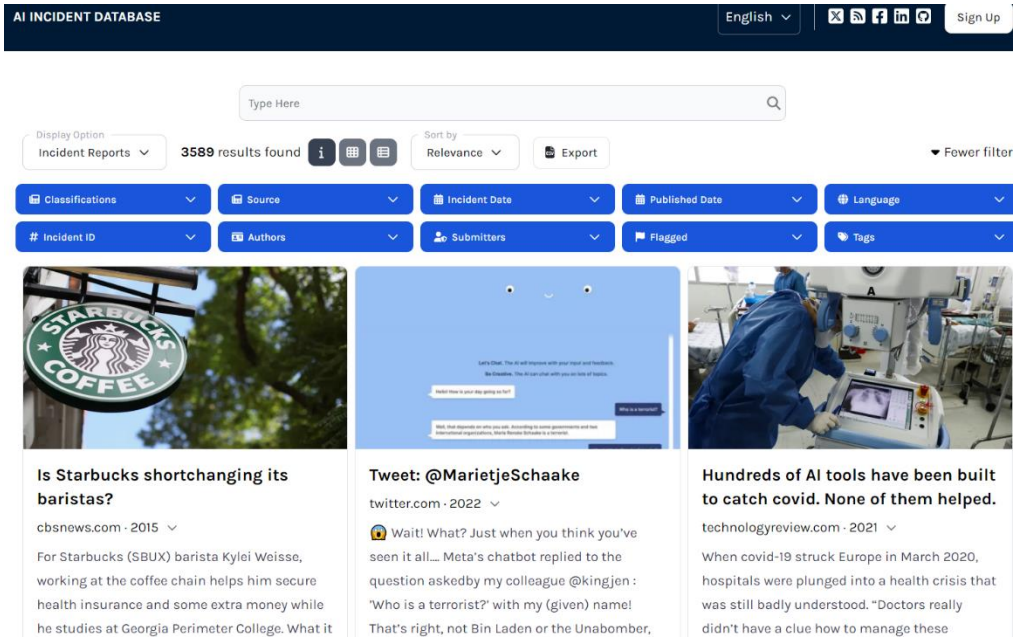


Figure 2. Screenshot of the AIID's '*discovery*' query tool

Since AIID did not provide a direct method for filtering incidents involving *workers*, we downloaded the whole dataset of 3,545 full-text news reports collected into 721 incidents as bulk data on August 1st, 2024. Initially, we explored the download by querying for each of the three keywords *labor*, *jobs*, and *worker*; we identified additional terminology from the incidents with these keywords in their associated news reports and ultimately constructed a dataset by searching for the following queries:

- *Compensation*
- *Firing*
- *Hiring*
- *Scheduling*
- *Unemployment*
- *Worker death*
- *Workplace*

We recorded the number of results for each query and deduplicated, retrieving 304 keyword matches in 266 news reports grouped into 126 incidents, which were associated with a total of 1,183 news reports, both retrieved by and not retrieved by keywords. We calculated the number of news reports retrieved by each keyword and grouped by incident.

Ultimately, we categorised an incident as labour related only when a news report retrieved by a keyword indicated how AI has already contributed to harms or risks for workers' well-being in the

workplace. Underlying our decision of whether an incident was eligible for inclusion in our labour-related dataset was the question: did automation or implementation of AI in the workplace either replace, disrupt, or augment the tasks and prospects of safe employment for workers? For instance, for celebrities and creative artists, we considered their capacity to extract value, so that we treated their bodies, personality, styles, or likeness as an intrinsic part of their labour.

We kept an inventory of thematic codes to classify recurring types of AI technologies, examples of workplace tasks, reported harms and risk mitigation topics included in the labour-related dataset. We excluded speculated harms that have not occurred, or that do not unambiguously bear directly on the workplace, such as gender or racial bias in search results for professions. Since we wanted to focus on AI's own impact on labour, we also excluded from the labour-related dataset:

1. Incidents related to the internal management of AI-focused companies (e.g., hiring, firing and leadership changes)
2. Incidents that originated outside employment, for instance related to:
    - Administration of social benefits, governance, justice, and law enforcement. Though relevant to labour, the administration of unemployment and social security benefits were considered to be outside the scope of the workplace.
    - Harmful representations pertaining to culture (e.g., generative AIs return violent imagery), content delivery and information retrieval (e.g., Image search for 'CEO' returns predominantly male-presenting results).
    - Anti-competition and monopolistic firm behaviour, except those specific to the job market
    - Deep fakes, except those generated as a condition of employment

Ultimately, we manually classified 57 incidents (with 184 keyword matches in 112 of their associated news reports – see Figure 3) as directly involving the effects of AI on labour. We compared these manually classified labour-related incidents with a keyword-based retrieval strategy: we collected the number of news reports retrieved by each keyword, grouped them by incident, and collected all news reports associated with the matching incidents.

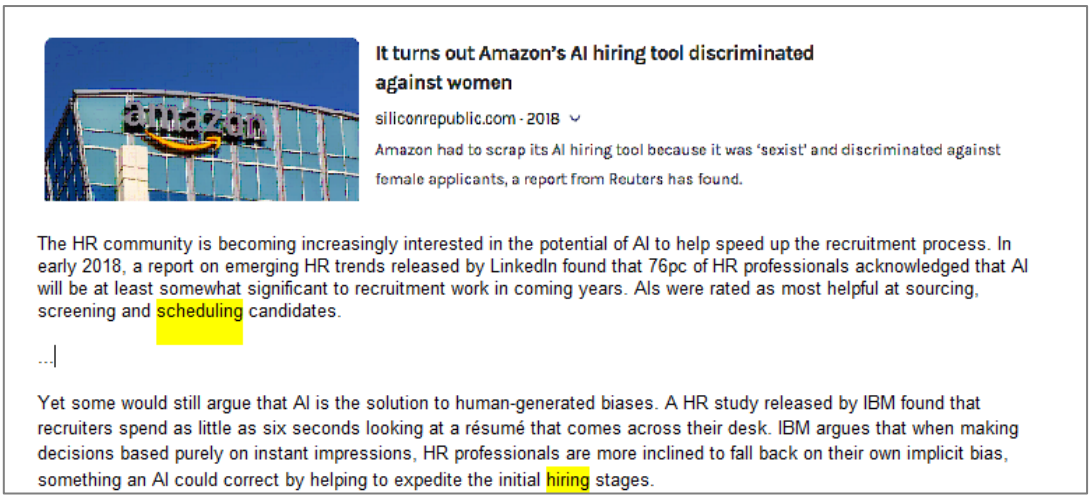


Figure 3. Sometimes multiple keyword searches retrieve a single news report: both '*scheduling*' and '*hiring*' are found in the news report (Short, 2018). The news report Short (2018) is associated with AIID's Incident 37 'female applicants down-ranked by Amazon recruiting tool.'

Results

AI incident tracker 1: OECD AIM

We compared the OECD AIM database as a whole to the subset of workers-related incidents (600 incidents with 6,744 associated news reports that the OECD AIM database categorised as having *workers* as '*affected stakeholders*'). Future threats are more pronounced for Workers: we found that 35% of *workers*-related incidents are marked as '*future threat*' compared to only 25% of incidents in the OECD AIM database as a whole. Figure 4 shows the percentage of all incidents involving workers broken down by industry sectors. The share of future threats with *workers* listed as affected as stakeholders (Figure 4 column 3) is particularly high for two sectors: *arts, entertainment, and recreation* (70%) and *business processes and support services* (49%) [Group 4]. Some industries [Group 1] have a larger percentage of future threat incidents but face little AI-related threat specifically to *workers*, now or in the future, such as *agriculture*; *energy, raw materials, and utilities*; *real estate*; and *environmental services*. Other sectors [Group 2] not only have incidents with minimal or zero AI impact for workers (Figure 4 column 2), but relatively few future threats generally (Figure 4 column 4): *food and beverages*; *construction and air conditioning*; *consumer products*; *consumers services*; *and travel, leisure, and hospitality*. Other industry sectors [Group 5] contain a moderate percentage of future threats, but their incidents do not proportionally affect *workers* more than other stakeholders (Figure 4 column 2); however, for this group, an incident is more likely to be a future threat when involving Workers as stakeholders in these industries, as seen in Table 1 column 1 for Group 5: *healthcare, drugs and biotechnology* (69%); *government, security and defence* (63%); *financial and insurance* (57%); *IT infrastructure and hosting* (58%); *logistics wholesale, retail* (50%); *robots, sensors, IT hardware* (53%). For some other industries [Group 3], *workers* have been identified as stakeholders in incidents, but incidents are less likely to be future threats, from Table 1 column 1: *digital security* (41%); *education and training* (24%); *media, social platforms, marking* (30%); *mobility* and autonomous vehicles (28%). These sectors face a moderate percentage of future threats that mostly concern other concerns besides *workers*, as seen in Figure 4 column 2 for Group 3.

Group	OECD Industry Sector	% of Stakeholder Worker Incidents that are Future Threat	% of Future Threat Incidents that are Stakeholder Worker	% All incidents that are Stakeholder Worker	% All Incidents that are Future Threat
1	Energy, raw materials and utilities	0%	0%	0%	60%
	Agriculture	0%	0%	0%	67%
	Environmental services	0%	0%	13%	88%
	Real estate	0%	0%	11%	33%
2	Travel, leisure and hospitality	0%	0%	0%	22%
	Construction and air conditioning	0%	0%	7%	7%
	Consumer products	0%	0%	4%	19%
	Consumer services	17%	17%	11%	11%
	Food and beverages	0%	0%	0%	0%
3	Digital security	41%	10%	9%	37%
	Education and training	24%	21%	33%	38%
	Media, social platforms, marketing	30%	20%	17%	25%
	Mobility and autonomous vehicles	28%	16%	15%	27%
4	Business processes and support services	46%	49%	39%	37%
	Arts, entertainment and recreation	35%	70%	44%	22%
5	Financial and insurance	57%	18%	13%	42%
	Government, security and defense	63%	12%	8%	40%
	Healthcare, drugs and biotechnology	69%	20%	13%	44%
	IT infrastructure and hosting	58%	31%	23%	44%
	Logistics wholesale, retail	50%	8%	18%	36%
	Robots, sensors, IT hardware	53%	29%	21%	39%

Less ← Future threat and/or risk exposure to workers → More

Figure 4. The proportion of OECD AIM incidents with workers as a stakeholder, broken down by each OECD-classified Industry, indicating the proportion that the database classified as '*future threat*'.

AI incident tracker 2: AIID

The 57 labour-related incidents were not the only incidents retrieved with the query terms we chose; Table 1 shows that different percentages of labour-related incidents were retrieved for different query terms, ranging widely, from 6/6 (100%) for scheduling to 11/33 (33%) for compensation. While we classified incidents as labour-related or not labour-related at the incident level, different articles related to the same incident presented the incident differently.

To examine the strength of the keyword-based signal that the incident should be interpreted as labour-related, we examined how many news reports associated with labour-related incidents were returned by a given keyword, also shown in Table 1.

		compensation	**firing**	**hiring**	**scheduling**	**unemployment**	**worker death**	**workplace**	**total matches**	**deduplicated**
Number of news reports retrieved by the keyword	From Labour-related Incidents	16	20	55	15	4	30	46	186	**157**
	From All Incidents	51	28	84	15	19	40	68	304	**266**
Number of matching incidents (news reports retrieved by the keyword, grouped by incident)	Labour-related Incidents	11	11	21	6	4	8	22	86	**57**
	All Incidents	33	19	44	6	8	17	37	164	**127**
Total news reports associated with matching incidents	From Labour-related Incidents	86	164	189	52	40	84	353	968	**533**
	From All Incidents	320	246	425	52	108	137	519	1807	**1183**

Table 1. The number of incidents and labour-related incidents returned for each query term in AIID. We deduplicated the total since the same news report may be retrieved by multiple keyword searches.

We iteratively classified AIID's labour-related AI incidents as shown in Table 2. This resulted in a total of 33 labels which we grouped into 4 categories: technology, workplace task, risks, and policy domain.

Technology	Workplace Task	Risks	Policy Domain
- Robotics - Algorithmic design - Autonomous driving - Predictive analytics - Generative AI - Computer vision - Networks - Manual data classification	- Warehouse operations - Recruitment, Personnel and hiring decisions - Assembly - Automating contracts and business rules - Security monitoring and surveillance - Termination decisions - Data classification - Job performance assessment - Journalism and reporting - Health and safety protocol - Creative content production - Content filtering - Jurisprudence - Chatbot - Conduct and behaviour	- Occupational hazard overwork or fatigue - Physical harm - No human override - Gender inequity - Racial inequity - Unreliable information - Psychological harm - Technological under comprehension - Human attribution issues - Financial harm or reputational harm - Political harm	- Labour regulations and workplace protections - Worker privacy, likeness, and labour ownership - Scaling automation versus labour cost benefit - Employment, contracting and termination - Unreasonable job expectations, assessment, and disciplining - Worker privacy, likeness, and labour ownership - Worker integrity and whistleblowing protections - Content moderation/Data labour

Table 2. Our own categorization of AIID's labour-related AI incidents into 33 labels in 4 categories

Discussion

The AI incident trackers do not facilitate ready retrieval of labour-related incidents. Being unable to readily see the incidents related to labour makes it impossible to understand what problems are associated with the AI-related risks and harms to labour.

The two AI incident trackers we examined used limited existing labour-related terminology. In OECD AIM, we found only one relevant term—the stakeholder *workers*—which is applied not only to AI end-users whose labour is replaced or augmented by AI systems, but also to AI producers (e.g., workers under pressure to engineer AI systems).

In AIID we identified three relevant term sets—'*data input*', '*AI task*' and '*end user amateur/expert*'—from the CSET taxonomy manual for incident classification. However, '*data input*' did not consider the human labour and roles involved in producing input data (such as training data that the AI system was trained on). '*AI task*' considers tasks such as '*resume reading*' or '*chatbot*', which can be viewed as labour replacement (Hoffmann et al., 2023). The inherent problem of staffing a support hotline for eating disorders with chatbots instead of a human call centre workforce, for example, is that this violates our expectations that therapeutic applications may require emotional labour (Posada, 2020), but '*AI task*' elides emotional labour.

While the OECD AIM has a taxonomy for industry sectors, its reliance on news reports risks overrepresenting some sectors such as entertainment. *Workers* stakeholder incidents tend to concentrate in the *arts, entertainment, and recreation* industry sector, covering workers' copyright infringement, data protections and personal likeness. And the challenges of generative AI in the workplace have been raised by recent Hollywood strikes calling for controls over licensing creative

content. More attention should be paid to understanding AI's effects on different sectors, especially given the likelihood that some sectors are overrepresented. It was unclear when reconciling how classification terms like worker and '*future threat*' should be interpreted when combined. Should policymakers rely on OECD AIM's classification technique when it excludes workers as stakeholders and simultaneously designates incidents in industrial sectors pertaining to earthborn land uses (like real estate, energy, agriculture, etc.) as '*future threat*?' Representing incidents in this manner presents its own risk and high stakes: Can policymakers properly gauge why heightened future risks in a sector should not portend in harm for workers?

The heavy reliance on news coverage, common to current AI trackers (Turri & Dzombak, 2023), leads to some challenges. News reports can frame different stories with the same evidence, by choosing what to ignore, background or highlight. When Amazon's implementation of a predictive hiring algorithm resulted in gender discrimination (Dastin, 2018), some news reports attributed the harm to sexist input data (Kraus, 2018) or mischaracterised output from reinforcement learning algorithms as outperforming human judgment (Short, 2018) and free from '*user abuse*' (Doctorow, 2018; Papadopaulos, 2018). Current AI incident report approaches group incidents geographically and temporally, which reduces the structured data available about where and when incidents took place. Consequently, in AIID, determining which incidents are labour related requires a substantial amount of judgement. When a single news report frames the news event as a labour issue, it makes the case that the incident is labour-related. Yet each incident collects multiple news reports, and labour was often in the background or used as supplemental frame. Finding incidents for labour-related policy topics was different from finding mentions of keywords. Some keywords have high precision, retrieving only labour-related incidents, but low recall, giving a limited picture of the risks and harms. For instance, scheduling only matched incidents about controversial algorithmically driven worker shift management software. Keywords with stronger recall, like '*compensation*' and '*hiring*', tended to be less precisely labour-focused, with matching incidents often referring to internal AI enterprise practices and consumer settlements.

Power relationships between the working class and management (and by extension, capital) are salient in the labour-related incidents we flagged: who introduces AI into the workplace? Who is responsible for AI when it backfires? Automated staffing decisions using AI-supported predictive analytics led to multiple problems (e.g., Southwest Airlines' flight cancellations (Sider, 2022), Tesla's factory delays (Duhigg, 2018)). Lack of human override and management's technical under-comprehension exacerbated these problems. Values need to be taken into consideration to determine which labour conflicts themselves qualify as harms.

Future work

Future work should examine how the distribution of news reports used in AI incident trackers vary over time, across industries and in relation to AI principles and news media sources and audiences. The ratio of news reports to incidents varies, with some incidents (such as the Hollywood strikes) heavily reported, without regard to the actual exposure to AI-related harms in a given industrial sector. Comparing how news reports frame the same AI incident (e.g., with framing analysis) would be valuable. Likewise, researchers should seek to understand how different groupings into incidents can contribute to policymakers' problem definitions, perhaps by examining the variation in how news reports are grouped in different AI incident trackers.

Future research could systematically identify which query terms to use, including testing stemming for words such as hiring, firing, and scheduling as well as considering additional terminology such as '*crowd labor*'. The use of terminology in AI governance could also be fruitfully examined, e.g., '*trustworthy* AI', '*ethical* AI', '*AI for good*', '*beneficial* AI', and '*responsible* AI' (Stix, 2022).

Current definitions for harm or hazard (Placani, 2017; Rowe, 2021) need to be revisited when evaluating the relationship between AI and labour to enable incident risk prevention (Meyer, 2023). Incident reporting, which has roots in risk management and safety (Johnson, 2003), may not be sufficient for understanding AI risks to workers. Alternative conceptual frameworks for risk management from epidemiology, audit culture and social work may be more suitable for centring workers.

Future work should investigate what harm, incident, potential harm, or hazard mean in the context of labour. For instance, are '*disruptions*' ever supposed to be weathered as part of events contributing to the normal functioning of the economy? Tradeoffs must consider multiple perspectives (e.g., worker, manager, capital) in the political economy of labour.

Sociotechnical perspectives will be needed. Above all, it is important to consider *'how algorithms may reshape organizational control'*, as Kellog et al. (2020) review. Sociotechnical analysis of fairness in machine learning (Selbst et al., 2019) can inspire new approaches for attending to the power dynamics of AI and labour, drawing on fields such as organizational behaviour and management science, human resources management, labour law and the sociology of labour.

Future classifications of AI and labour should document when a system extracts labour value unjustly. Key examples would be when a '*worker*' can claim direct credit or likeness to the source data (such as Scarlett Johanssen's voice appropriated by OpenAI (Pisani & Albert, 2024)) or when input data is created with the express purpose of serving an AI system (Satariano & Mozur, 2023) (see our content moderation/data labour category in Table 3). Taxonomies for organizing how AI interacts with prior types of '*people problems*' inside the workplace (Moore, 2019) may be helpful since both AI and AI incident reporting ultimately depend on how people are organised to share information within a firm. Standardised terminologies and definitions for workplace safety can be informed by International Labour Organization worker protections requiring information disclosure in the event of mass dismissal and redundancy (De Stefano, 2019). Incidents are made visible based on the power structures upheld by a given taxonomy: naming and power is worth specific consideration.

Conclusions

The two incident trackers we investigated do not adequately capture the nuanced impacts of AI on labour. Particular attention needs to be paid to power imbalances that may increase risks for harm in the workplace. Better definitions are needed to capture labour-related AI incidents, to help policymakers gather evidence to anticipate and mitigate the risks and harms threatening workers across diverse industries. AI incident trackers' reliance on news reports and limited vocabulary for identifying worker-related harms and risks leads to gaps in understanding AI for policy formulation and problem definition.

Are the labour-related risks of AI intelligible in the incident trackers with a level of detail and reliability that could support comprehensive policy and problem definition for workers' protection from the harms of AI? Our answer is no.

Acknowledgements

This research is funded by the United States Institute of Museum and Library Services RE-250162-OLS-21. Thanks to Zachary Kilhoffer for discussions of AI trackers and to Corinne McCumber, Heng Zheng, Yuanxi Fu and Zhixuan 'Kyrie' Zhou for providing feedback on a draft. Jodi Schneider was supported in part as the 2024–2025 Perrin Moorhead Grayson and Bruns Grayson Fellow, Harvard Radcliffe institute for advanced study. CRediT roles: TL: conceptualization, data curation, formal analysis, methodology, visualization, writing – original draft, writing – review and editing; Jodi Schneider: conceptualization, funding acquisition, supervision, writing – review and editing.

About the author

Theodore Dreyfus Ledford is a PhD student in Information Sciences at the University of Illinois at Urbana-Champaign. tledfo2@illinois.edu

References

Abercrombie, G., Benbouzid, D., Giudici, P., Golpayegani, D., Hernandez, J., Noro, P., Pandit, H., Paraschou, E., Pownall, C., Prajapati, J., Sayre, M. A., Sengupta, U., Suriyawongkul, A., Thelot, R., Vei, S., & Waltersdorfer, L. (2024). *A collaborative, human-centred taxonomy of AI, algorithmic, and automation harms* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.01294

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2), 3–30. https://doi.org/10.1257/jep.33.2.3

Arda, S. (2024). *Taxonomy to regulation: A (geo)political taxonomy for AI risks and regulatory measures in the EU AI act* [Preprint]. arXiv. http://arxiv.org/abs/2404.11476

Arntz, M., Gregory, T., & Zierahn, U. (2016). The risk of automation for jobs in OECD countries: A comparative analysis (OECD Social, Employment and Migration Working Papers, 189). OECD. https://doi.org/10.1787/5jlz9h56dvq7-en

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. Journal of Economic Perspectives, 29(3), 3–30. https://doi.org/10.1257/jep.29.3.3

Autor, D.H. (2022). The labor market impacts of technological change: From unbridled enthusiasm to qualified optimism to vast uncertainty (Working Paper w30074). National Bureau of Economic Research. https://doi.org/10.3386/w30074

Baumgartner, F. R., & Jones, B. D. (2015). The politics of information: Problem definition and the course of public policy in America. University of Chicago Press.

Booth, H., Rike, D., & Witte, G. A. (2013 December). The National Vulnerability Database (NVD): Overview (Information Technology Laboratory Bulletin Series). National Institute of Standards. https://www.nist.gov/publications/national-vulnerability-database-nvd-overview

Cameron, J. (Director). (1984, October 26). The Terminator [Film]. Cinema ’84; Euro Film Funding; Hemdale.

Cattell, S., Ghosh, A., & Kaffee, L.-A. (2024). Coordinated flaw disclosure for AI: Beyond security vulnerabilities [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2402.07039

Critch, A., & Russell, S. (2023). TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2306.06924

Dastin, J. (2018, October 10). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

De Stefano, Valerio. (2019). ‘Negotiating the algorithm’: Automation, artificial intelligence, and labor protection. Comparative Labor Law & Policy Journal, 41(1), 15-46.

Doctorow, C. (2018, October 11). Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist. Boing Boing. https://boingboing.net/2018/10/11/garbage-conclusions-out.html

Duhigg, C. (2018, December 13). Dr. Elon & Mr. Musk: Life inside Tesla's production hell. Wired. https://www.wired.com/story/elon-musk-tesla-life-inside-gigafactory/

Head, B. W. (2010). Reconsidering evidence-based policy: Key issues and challenges. Policy and Society, 29(2), 77–94. https://doi.org/10.1016/j.polsoc.2010.03.001

Hoffmann, M., & Frase, H. (2023). Adding structure to AI harm. Center for Security and Emerging Technology. https://doi.org/10.51593/20230022

Hoffmann, M., Narayanan, M., Mitra, A., Liao, Y.-J., & Frase, H. (2023). CSET AI Harm Taxonomy for AIID and Annotation Guide. https://github.com/georgetown-cset/CSET-AIID-harm-taxonomy

Hutiri, W., Papakyriakopoulos, O., & Xiang, A. (2024). Not my voice! A taxonomy of ethical and safety harms of speech generators. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 359–376. https://doi.org/10.1145/3630106.3658911

Johnson, C. (2003). Failure in safety critical systems: A handbook of accident and incident reporting. University of Glasgow Press.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.). (2000). Creating safety systems in health care organizations. In To err is human: Building a safer health system (pp. 155–204). National Academies Press. https://doi.org/10.17226/9728

Kraus, R. (2018, October 10). Amazon's sexist recruiting algorithm reflects a larger gender bias. Mashable. https://mashable.com/article/amazon-sexist-recruiting-algorithm-gender-bias-ai

Laker, B. (2023, October 30). The future of work: Navigating the complex landscape of flexibility. Forbes. https://www.forbes.com/sites/benjaminlaker/2023/10/30/the-future-of-work-navigating-the-complex-landscape-of-flexibility/

Lane, M., & Saint-Martin, A. (2021). The impact of Artificial Intelligence on the labour market: What do we know so far? (OECD Social, Employment and Migration Working Papers 256). https://doi.org/10.1787/7c895724-en

MacKillop, E., & Downe, J. (2023). What counts as evidence for policy? An analysis of policy actors' perceptions. Public Administration Review, 83(5), 1037–1050. https://doi.org/10.1111/puar.13567

Macrae, C. (2016). The problem with incident reporting. BMJ Quality & Safety, 25(2), 71–75. https://doi.org/10.1136/bmjqs-2015-004732

McGregor, S. (2021). Preventing repeated real world AI failures by cataloguing incidents: The AI incident database. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817

Meyer, C. O. (2024). Can one 'prove' that a harmful event was preventable? Conceptualizing and addressing epistemological puzzles in post incident reviews and investigations. Risk, Hazards & Crisis in Public Policy, 15(3), 374–392. https://doi.org/10.1002/rhc3.12281

Moore, P. V. (2019). The mirror for (artificial) intelligence: In whose reflection? Automation, artificial intelligence, & labor law. Comparative Labor Law & Policy Journal, 41(1), 47–68.

NASA Aviation Safety Reporting System. (n.d.). ASRS: the case for confidential incident reporting systems (ASRS Research Papers 60). NASA Aviation Safety Reporting System. https://asrs.arc.nasa.gov/docs/rs/60_Case_for_Confidential_Incident_Reporting.pdf

OECD. (2022a). OECD framework for the classification of AI systems (OECD Digital Economy Papers 323). OECD. https://doi.org/10.1787/cb6d9eca-en

OECD. (2022b). Harnessing the power of AI and emerging technologies: Background paper for the CDEP Ministerial meeting (OECD Digital Economy Papers 340). OECD. https://doi.org/10.1787/f94df8ec-en

OECD. (2023). Stocktaking for the development of an AI incident definition (OECD Artificial Intelligence Papers 4). OECD. https://doi.org/10.1787/c323ac71-en

OECD. (2024). Defining AI incidents and related terms (OECD Artificial Intelligence Papers 16). OECD. https://doi.org/10.1787/d1a8d965-en

Papadopoulos, L. (2018, Oct 12). Amazon shuts down secret AI recruiting tool that taught itself to be sexist. Interesting Engineering. https://interestingengineering.com/innovation/amazon-shuts-down-secret-ai-recruiting-tool-that-taught-itself-to-be-sexist

Pawson, R., Wong, G., & Owen, L. (2011). Known knowns, known unknowns, unknown unknowns: The predicament of evidence-based policy. American Journal of Evaluation, 32(4), 518–546. https://doi.org/10.1177/1098214011403831

Pisani, J., & Albert, V. (2024, May 20). Scarlett Johansson rebukes OpenAI over 'Eerily Similar' ChatGPT voice. Wall Street Journal. https://www.wsj.com/tech/ai/openai-chatgpt-sky-voice-scarlett-johansson-43d13bbf

Placani, A. (2017). When the risk of harm harms. Law and Philosophy, 36, 77–100. https://doi.org/10.1007/s10982-016-9277-x

Posada, J. (2020). The future of work is here: Toward a comprehensive approach to Artificial Intelligence and labour [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2007.05843

Responsible AI Collaborative. (2022). Founding Report. Responsible AI Collaborative. https://asset.cloudinary.com/pai/cf01cce1af65f5fbb3d71fa092d001db

Reynard, W. D. (1986). The development of the NASA aviation safety reporting system (NASA reference publication 114). National Aeronautics and Space Administration, Scientific and Technical Information Branch. Available as ASRS Research Papers 34 https://asrs.arc.nasa.gov/docs/rs/34_Development_of_NASA_ASRS.pdf

Rodrigues, R., Resseguier, A., & Santiago, N. (2023). When Artificial Intelligence fails: The emerging role of incident databases. Public Governance, Administration and Finances Law Review, 8(2), 17–28. https://doi.org/10.53116/pgaflr.7030

Rowe, T. (2021). Can a risk of harm itself be a harm? Analysis, 81(4), 694–701. https://doi.org/10.1093/analys/anab033

Satariano, A., & Mozur, P. (2023, February 7). The people onscreen are fake. The disinformation is real. The New York Times. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

Schiff, D. S. (2024). Framing contestation and public influence on policymakers: Evidence from US artificial intelligence policy discourse. Policy and Society, 43(3), 255–288. https://doi.org/10.1093/polsoc/puae007

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598

Shelby, R., Rismani, S., Henne, K., Moon, AJ., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 723–741. https://doi.org/10.1145/3600211.3604673

Short, E. (2018, October 11). It turns out Amazon's AI hiring tool discriminated against women. Silicon Republic. https://www.siliconrepublic.com/careers/amazon-ai-hiring-tool-women-discrimination

Shrishak, K. (2023). How to deal with an AI near-miss: Look to the skies. Bulletin of the Atomic Scientists, 79(3), 166–169. https://doi.org/10.1080/00963402.2023.2199580

Sider, A. (2022, December 28). How Southwest Airlines melted down. Wall Street Journal. https://www.wsj.com/articles/southwest-airlines-melting-down-flights-cancelled-11672257523

Stix, C. (2022). Artificial intelligence by any other name: A brief history of the conceptualization of "trustworthy artificial intelligence." Discover Artificial Intelligence, 2, 26. https://doi.org/10.1007/s44163-022-00041-5

Turri, V., & Dzombak, R. (2023). Why we need to know more: Exploring the state of AI incident documentation practices. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 576–583. https://doi.org/10.1145/3600211.3604700

van der Kleij, R., Schraagen, J. M., Cadet, B., & Young, H. (2022). Developing decision support for cybersecurity threat and incident managers. Computers & Security, 113, 102535. https://doi.org/10.1016/j.cose.2021.102535

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 214–229. https://doi.org/10.1145/3531146.3533088

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Appendix

Data availability

Data underlying this research is available at https://doi.org/10.13012/B2IDB-1156758_V1

Information Research – Vol. 30 No. iConf (2025)

Cultivating university data culture in the age of artificial intelligence: a conceptual framework and critical reflections

Yunjie Tang

DOI: https://doi.org/10.47989/ir30iConf47293

Abstract

Introduction. This paper examines the transformation of university data culture in the age of AI. AI and big data are reshaping how universities manage data, make decisions, and interact with stakeholders, raising both opportunities and challenges.

Method. The study explores the concept of university data culture, proposing a four-layer conceptual framework: the data layer, user layer, organizational layer, and societal layer.

Analysis. The four-layer framework captures the lifecycle of data, stakeholder behaviours and competencies, institutional policies, and societal influences. It emphasizes how AI transforms data usage and governance within universities, as well as the ethical and social considerations that arise.

Results. AI is driving a shift from traditional data management to structured data governance, fostering greater data-informed decision-making, and encouraging bottom-up participation among faculty and students. However, challenges related to ethics, privacy, and inequality persist.

Conclusion. The transformation of university data culture is not just technological but cultural, requiring universities to address ethical and social issues. Ensuring inclusivity, fairness, and responsibility in data practices is critical for fostering a balanced and sustainable data culture in higher education.

Introduction

Mayer-Schönberger (2013) describes the big data era as '*a radical revolution in how we live, work, and think*', with big data and artificial intelligence (AI) increasingly becoming fundamental to human civilization and shaping a new cultural paradigm (Shi & Feng, 2017). As key drivers of knowledge production and dissemination, universities are now entering a data-driven era, harnessing these technologies to improve teaching, research, and administrative processes. This shift has given rise to a new data culture within higher education, which is transforming management practices, pedagogical approaches, and research methodologies. Universities are not only recognizing the practical uses of data but also its potential to fundamentally reshape organizational structures and decision-making processes. However, the integration of AI presents both opportunities and challenges, as it redefines how data is understood, analyzed, and applied.

This paper uses content analysis to investigate the evolution of university data culture in the AI era, focusing on the core components of this cultural shift and AI's role in fostering it. Further, it draws on Bourdieu's concept of cultural capital to critically examine the ethical, privacy, and social responsibility issues that arise, offering strategies to address these challenges.

From data culture to university data culture

As a relatively new and evolving concept, data culture lacks a clear consensus across academic and industrial communities. Different interpretations arise based on varying research objectives and contexts. Broadly, data culture can be seen through several lenses: as part of societal culture focused on data-driven (Zhang & Lu, 2020), logical decision-making (cultural subordination); as a key component of data competency (Kee-Luen et al., 2018), emphasizing evidence-based decision-making for value creation (data competency); and as a set of organizational practices and values related to data use, sharing, and management (Integration) (Liu & Tu, 2023; Oliver et al., 2023; Oliver et al., 2024). Oliver et al. (2024) proposed five core aspects of data culture: data-related skills and attitudes (such as literacy and fluency), data sharing behaviours, data reuse, data ethics and governance, and localized perspectives like data sovereignty. These components reflect the multifaceted nature of data culture, which operates across various societal, technical, and organizational levels.

Data culture in higher education has been defined in various ways. The Institute of Education Sciences views it as an institutional environment shaped by attitudes, values, and practices around data use, reflecting leadership's vision (Hamilton et al., 2009). Powers and Henderson (2018) describe it as a collective consciousness that emphasizes consistency and reciprocity in data practices, aiming to democratize data for actionable insights that support the university's mission. Similarly, Raffaghelli et al. (2020) highlight the contextual nature of data culture, shaped by institutional traditions and roles, and emphasize that it can be transformed by stakeholders to promote equity.

Drawing from these perspectives, university data culture can be defined as a collective awareness and institutional framework governing how data is processed, used, and managed. It encompasses the values, policies, and behaviors of administrators, faculty, and students, driving data-informed decision-making. University data culture emphasizes data democratization, ensuring broad access and transparency, and transforming data into actionable insights that support teaching, research, and administration. This shift is not only technological but cultural, fostering shared data practices across institutional roles.

Furthermore, university data culture is shaped by each institution's socio-historical context. It requires critical reflection on ethical and social implications, such as privacy and fairness. By addressing these concerns, universities can reshape data practices to create more inclusive environments. Ultimately, university data culture drives intelligent transformation, enabling

universities to leverage AI and big data to enhance learning, research, and decision-making while adapting to evolving societal needs.

Proposed conceptual framework of university data culture

Powers and Henderson (2018) identify four core characteristics of university data culture: sharing data, valuing data, trusting data, and using data. These features provide a foundational framework for understanding the role of data in universities. However, university data culture extends beyond these traits, involving deeper institutional arrangements, technological infrastructure, and ethical considerations. To offer a more comprehensive analysis, this study builds upon the three-tier information culture framework proposed by Oliver and Foscarini (2021) and, in conjunction with the data culture dimensions identified by Oliver et al. (2024), develops an initial four-layer conceptual framework, as shown in Figure 1. The focus of each layer is briefly outlined as follows:

- **Data layer:** Emphasises the flow of information, including how data is created, preserved, and managed. This layer focuses on the technical and procedural aspects of data handling, ensuring that data lifecycle processes are efficient and effective.
- **User layer:** Concentrates on users' data literacy and behavioural norms, focusing on individuals' various behaviours in data practices. This includes enhancing data competencies among faculty, students, and administrators to foster a data-informed community.
- **Organizational layer:** Focuses on elements that are less apparent yet still identifiable, involving data-related regulations, policies, systems, and infrastructures. This layer addresses the formal structures and guidelines that govern data management and governance within the university.
- **Societal layer:** Addresses more macro and the most elusive aspects, rooted in the cultural beliefs embedded in the collective and individual subconscious. This is the deepest and most subtly perceived cultural level, encompassing personal and collective unconscious beliefs, attitudes, and ideologies about information, as well as data governance philosophies. It includes the ethical frameworks and societal norms that influence how data is perceived and utilised within the institution.

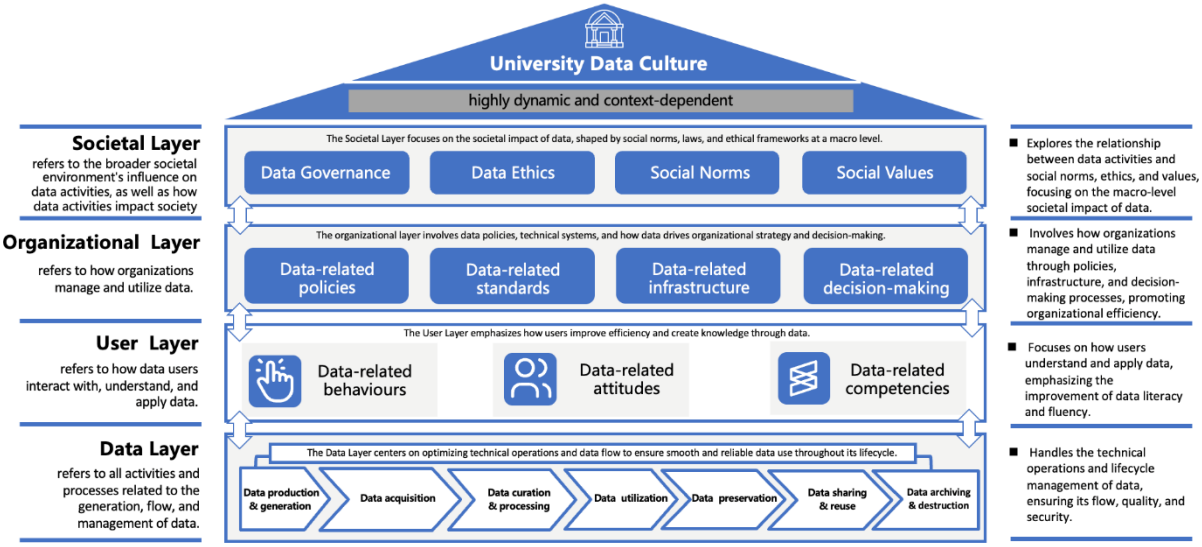


Figure 1. The proposes a four-layer conceptual framework

The data layer centers on the lifecycle of data within universities, encompassing its creation, collection, processing, management, sharing, and reuse. In higher education, data has emerged as a critical asset, with universities generating vast amounts of information from student learning behaviors, research activities, and administrative operations. To manage this data, many institutions have developed data management platforms and '*data lakes*' that facilitate the storage and analysis of diverse data sources (Villegas-Ch et al., 2018). For example, student performance data can be consolidated to build early warning systems, allowing administrators to identify at-risk students and offer personalized academic support. This layer emphasizes optimizing data flow and ensuring its effective integration across various functions to enhance institutional performance.

The user layer focuses on the behaviours, attitudes, and competencies of key stakeholders, including faculty, students, and administrators in relation to data. Central to this layer are data literacy and fluency, which refer to users' ability to comprehend and effectively apply data in their teaching, research, and decision-making processes. Advanced users leverage data-sharing platforms and analytical tools to foster interdisciplinary research, driving innovation and promoting transparency across academic fields. This layer highlights the agency of users in shaping data usage within the university, as their familiarity with and engagement in data practices directly influence the institution's overall data culture.

The organizational layer addresses the role of universities as institutions, focusing on the policies, standards, infrastructure, and leadership that govern data usage. The development of a robust data culture within universities relies on clear policies that regulate the collection, storage, and use of data. For example, many universities implement stringent privacy protocols for learning management systems while simultaneously encouraging data sharing in research to promote collaboration. Leadership is a vital component of this layer; when university leaders advocate for data-driven decision-making and foster a supportive environment for data use, they significantly strengthen the institution's data culture (Powers & Henderson, 2018). This layer underscores the importance of organizational structure and governance in managing data effectively and ethically.

The societal layer considers broader societal factors, including data ethics, governance, and the influence of local and global contexts on university data culture (Deja, 2024). In the era of global data flows, universities must navigate complex issues such as data sovereignty (Lilley et al., 2024), privacy protection, and compliance with both local and international legal frameworks. For example, institutions engaged in cross-border research collaborations face challenges related to data sharing while adhering to national security concerns and privacy laws (Tang, 2024). This layer emphasizes the need to balance global data collaboration with respect for local legal, cultural, and ethical standards, ensuring that universities remain accountable to both societal and institutional norms.

Notably, university data culture is highly dynamic and context-dependent. Its form and function evolve with technological advancements, societal demands, and institutional developments. Different universities face unique challenges in constructing their data cultures: while technologically advanced institutions may more easily build comprehensive data cultures, those with limited resources may encounter greater technical and financial barriers. Therefore, building a data culture must account for specific contextual needs and constraints.

Transformation of university data culture in the AI era

The rise of AI is driving a transformation in university data culture that goes far beyond technical upgrades—it signals a fundamental shift in mindset. Universities are being called to undergo comprehensive, multidimensional changes to cultivate a data culture that aligns with their long-term goals. This transformation is not just about implementing new technologies; it requires creating an integrated environment rooted in collaboration and shared practices while also

reshaping organizational structures, management approaches, and cultural attitudes (Gagliardi et al., 2018).

One of the key shifts involves moving from '*data management*' to '*data governance*'. Traditional data management focuses primarily on storage and basic planning, but data governance requires a more structured approach that emphasizes role clarity, accountability, and institutionalized workflows. A shared governance system that involves all stakeholders is essential. For instance, universities have begun forming governance committees that integrate research and IT departments to oversee data sharing and ensure security and effective resource use.

Equally important is the transition from a '*reporting culture*' to a '*data-informed culture*'. In many institutions, data is still passively collected for performance reports, offering little value beyond compliance. A true data culture, however, sees faculty and administrators actively using data to drive decision-making and continuous improvement. A notable example is Georgia State University, which uses data analytics to identify academic risks and provide personalized support, significantly improving graduation rates.

Moreover, universities need to shift from a '*top-down*' to a '*bottom-up approach*'. While leadership support remains critical, broad participation by faculty and students is equally important to deeply embed a data-driven culture. One way to encourage this is by providing open-access data platforms and labs where faculty and students can engage in data analysis, fostering a grassroots data culture. This grassroots involvement can be a powerful catalyst for change (Zimpher, 2014).

The transition from '*single-driven*' to '*multi-stakeholder collaboration*' is crucial. Building a robust data culture requires partnerships not only within the institution but also with government, industry, and other external entities (Zhou, 2024). European universities, for example, have collaborated with tech companies to develop data tools that accelerate digital transformation and foster deeper integration with societal needs.

Furthermore, the focus should shift from simply building '*data infrastructure*' to creating a '*data ecosystem*'. While past efforts have centered on building data centers and servers, the rise of AI and big data calls for an ecosystem that facilitates data sharing and interdisciplinary collaboration (Vicente-Saez et al., 2020).

Finally, universities must evolve from '*traditional hierarchies*' to '*modern mechanisms*.' Bureaucratic structures are often too rigid to keep pace with the fast-evolving digital landscape. To adapt, universities should consider adopting flatter, more agile organizational models, such as establishing digital transformation offices or appointing chief digital officers (Davison et al., 2023).

Reflection: Data as a '*double-edged sword*' of new cultural capital

University data culture has undoubtedly fueled advancements in technology, management, and institutional practices. However, alongside this progress come ethical dilemmas, privacy concerns, and questions of social responsibility. Through the lens of Bourdieu's cultural capital theory, we see that when data transforms into a new form of capital, it deepens issues of inequality, power dynamics, and social reproduction (Grenfell, 2014; Ignatow & Robinson, 2017).

In Bourdieu's framework, cultural capital is a key determinant of power, and data, as a modern embodiment of this capital (Paino & Renzulli, 2013), plays a pivotal role in shaping who holds influence within universities. Those with the tools and expertise to manage and analyze data—administrators, researchers, and data scientists—naturally accumulate more capital, gaining leverage over decision-making processes. This shift in power risks exacerbating inequalities, as those without the requisite data skills become increasingly marginalized.

Moreover, Bourdieu emphasizes how cultural capital is reproduced within educational systems (Bourdieu, 1977). In today's data-driven environment, proficiency with data analysis becomes a gateway to accumulating this capital. Mastery of data analysis has become a key factor in gaining influence, with individuals and groups proficient in these skills often occupying leadership roles. Meanwhile, those lacking these competencies may find themselves marginalized. This imbalance not only affects university governance but also impacts students' learning opportunities and career prospects. The '*data divide*' between well-funded and under-resourced institutions further amplifies these discrepancies, limiting equitable access to advanced technologies.

The question of who controls data also reflects Bourdieu's ideas about power and ownership. In the realm of data culture, privacy concerns intensify as vast amounts of personal data, including student performance, research findings, and health metrics, are collected and analysed. Those who manage this data hold significant power, influencing behaviour through surveillance and evaluation. If left unchecked, biased algorithms used in processes such as admissions or resource allocation could perpetuate existing inequalities, further reinforcing the power structures they are intended to address.

Bourdieu's work also reminds us that universities, beyond being centres of knowledge production, are instrumental in the reproduction of social structures. In this context, the growing reliance on data culture, with its focus on quantifiable metrics, risks overshadowing more humanistic values. As institutions increasingly prioritize data-driven decision-making, emphasizing research outputs and student performance scores, there is a real danger that critical thinking, empathy, and moral responsibility may be neglected, ultimately undermining the broader mission of education to cultivate responsible global citizens.

To address these challenges, universities must adopt strategies that align with Bourdieu's warnings about the unequal distribution of capital. Enhancing data literacy is crucial to ensure that students and staff can equitably access and utilize data, while also understanding its ethical implications. Strengthening data ethics education is equally important, ensuring that individuals reflect on the power dynamics and privacy concerns inherent in data usage. Furthermore, promoting fair data usage through transparent policies will help democratize data access and reduce power imbalances. Finally, universities should adopt a diversified evaluation system that values ethical awareness, social responsibility, and critical thinking, alongside quantitative metrics.

Conclusion

The transformation of university data culture in the age of AI brings forward a conceptual framework that identifies the critical layers of data culture: data, user behaviour, organizational structures, and societal influences. AI plays a pivotal role in reshaping each of these components, fundamentally altering how universities manage data, make decisions, and engage with stakeholders. This shift is more than just a technological upgrade; it represents a deep cultural transformation within higher education. However, alongside these advancements come significant ethical, social, and operational challenges. Addressing these issues is essential to ensure that the evolving data culture within universities fosters inclusivity, fairness, and responsibility in its practices.

Given that this study is in its preliminary stages, the current framework was developed through a manual review of selected key literature related to data culture and information culture. This approach provided an initial foundation for understanding the multifaceted nature of university data culture. Moving forward, the research will adopt a more rigorous content analysis methodology, utilising qualitative analysis software such as NVivo, and will involve systematic, multi-level coding of relevant literature to enhance the scientific robustness and validity of the conceptual framework. Future research will ensure a comprehensive and empirically grounded exploration of the layers that constitute university data culture in the era of artificial intelligence.

About the author

Yunjie Tang is a Ph.D. candidate in the Department of Information Management at Peking University. An educational foundation has been established with a master's degree in Information Science and a bachelor's degree in Archival Science from Shanghai University. Current research is focused on data culture, governance, and privacy protection within a data-driven society. Publications have been featured in journals, including LISR, LIBRI, serials review, and library and information service. An upcoming book titled 'mechanisms for open sharing of scientific data' is set to be published. Correspondence may be directed to yunjietang@stu.pku.edu.cn.

References

Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge University Press.

Davison, R. M., Wong, L. H. M., & Peng, J. H. (2023). The art of digital transformation as crafted by a chief digital officer. *International Journal of Information Management*, 69, Article 102617. https://doi.org/10.1016/j.ijinfomgt.2022.102617

Deja, M. (2024). Information culture of university administration: making personnel bureaucracy a professional bureaucracy. Journal of Librarianship and Information Science, 56(2), 379-396. https://doi.org//10.1177/09610006221145931

Gagliardi, J., Parnell, A., & Carpenter-Hubin, J. (2018). The analytics revolution in higher education. *Change: The Magazine of Higher Learning*, 50(2), 22-29. https://doi.org//10.1080/00091383.2018.1483174

Grenfell, M. (2014). *Pierre Bourdieu: key concepts*. Routledge.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., Pickens, C., Martin, E., & Steele, J. L. (2009). *Using student achievement data to support instructional decision making*. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/dddm_pg_092909.pdf

Ignatow, G., & Robinson, L. (2017). Pierre Bourdieu: theorizing the digital. *Information, Communication & Society*, 20(7), 950-966. https://doi.org/10.1080/1369118X.2017.1301519

Kee-Luen, W., Hong-Leong, J. T., Shwu-Shing, N., & Min-Hooi, C. (2018). Insights from the CGMA Data Competencies Model: the role of data culture to the value creation process. *Journal of Telecommunication, Electronic and Computer Engineering*, 10(2-4), 187-192. https://jtec.utem.edu.my/jtec/article/view/4404

Lilley, S., Oliver, G., Cranefield, J., & Lewellen, M. (2024). Māori data sovereignty: contributions to data cultures in the government sector in New Zealand. *Information, Communication & Society*, 1-16. https://doi.org/10.1080/1369118X.2024.2302987

Liu, X., & Tu, Z. (2023). Thinking and exploring of data culture in the academic library. *Journal of Academic Libraries*, 41(6), 19-25. https://doi.org/10.16603/j.issn1002-1027.2023.06.003.

Mayer-Schönberger, V. (2013). *Big data: a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Oliver, G., Cranefield, J., Lilley, S., & Lewellen, M. (2023). Data cultures: a scoping literature review. *Information Research an international electronic journal*, 28(1), 3-29. https://doi.org/10.47989/irpaper950

Oliver, G., Cranefield, J., Lilley, S., & Lewellen, M. J. (2024). Understanding data culture/s: influences, activities, and initiatives: an annual review of information science and technology

(ARIST) paper. *Journal of the Association for Information Science and Technology*, 75(3), 201-214. https://doi.org/10.1002/asi.24737

Oliver, G., & Foscarini, F. (2021). *Recordkeeping cultures*. Facet.

Paino, M., & Renzulli, L. A. (2013). Digital dimension of cultural capital: the (in) visible advantages for students who exhibit computer skills. *Sociology of education*, 86(2), 124-138. https://doi.org/10.1177/0038040712456555

Powers, K., & Henderson, A. E. (2018). *Cultivating a data culture in higher education*. Routledge New York.

Raffaghelli, J. E., Manca, S., Stewart, B., Prinsloo, P., & Sangrà, A. (2020). Supporting the development of critical data literacies in higher education: building blocks for fair data cultures in society. *International Journal of Educational Technology in Higher Education*, 17, 1-22. https://doi.org/10.1186/s41239-020-00235-w

Shi, C., & Feng, X. (2017). From data to insights: cultural innovation and data science – a summary of the international academic conference on the national forum on cultural innovation and data science. *Journalism Research*, 33(5), 137-144.

Tang, Y. (2024). Privacy protection framework for open data: constructing and assessing an effective approach. *Library & Information Science Research*, 46(3), 101312. https://doi.org/10.1016/j.lisr.2024.101312

Vicente-Saez, R., Gustafsson, R., & Van den Brande, L. (2020). The dawn of an open exploration era: emergent principles and practices of open science and innovation of university research teams in a digital world. *Technological Forecasting and Social Change*, 156, Article 120037. https://doi.org/10.1016/j.techfore.2020.120037

Villegas-Ch, W., Luján-Mora, S., Buenaño-Fernandez, D., & Palacios-Pacheco, X. (2018). Big data, the next step in the evolution of educational data analysis. In *Proceedings of the International Conference on Information Technology & Systems* (ICITS 2018) (pp. 138-147). Springer International Publishing. https://doi.org/10.1007/978-3-319-73450-7_14

Zhang, J., & Lu, M. (2020). The status quo and cultivation of data culture in colleges and universities of China: on the role of institutional research. *Journal of Higher Education*, 41(12), 40-47. https://doi.org/1000-4203(2020)12-0040-08

Zhou, J. (2024). On the value, deviation, and orientation of digital transformation in double first-class universities. *Journal of Graduate Education*(3), 27-33. https://doi.org/10.19834/j.cnki.yjsjy2011.2024.03.05

Zimpher, N. L. (2014). *Building a smarter university: big data, innovation, and analytics*. State University of New York Press.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

'Sora is incredible and scary': public perceptions and governance challenges of text-to-video generative AI models

Kyrie Zhixuan Zhou, Abhinav Choudhry, Ece Gumusel, and Madelyn Rose Sanfilippo

DOI: https://doi.org/10.47989/ir30iConf47290

Abstract

Introduction. Text-to-video generative AI models such as Sora OpenAI have the potential to disrupt multiple industries. In this paper, we report a qualitative social media analysis aiming to uncover people's perceived impact of and concerns about Sora's integration.

Method. We collected and analysed comments (N=292) under popular posts about (1) Sora generated videos, (2) a comparison between Sora videos and Midjourney images, and (3) artists' complaints about copyright infringement by Generative AI.

Analysis. We employed a thematic analysis to analyse the collected comments.

Results. We found that people were most concerned about Sora's impact on content creation-related industries. Governance challenges included the for-profit nature of OpenAI, the blurred boundaries between real and fake content, human autonomy, data privacy, copyright issues, and environmental impact. Potential regulatory solutions proposed by people included law-enforced labelling of AI content and AI literacy education for the public.

Conclusions. Based on the findings, we discuss the importance of gauging people's tech perceptions early and propose policy recommendations to regulate Sora before its public release.

Introduction

As artificial intelligence (AI) continues to permeate daily life, advanced AI systems raise significant questions and prompt diverse social reactions. In particular, the remarkable advancements of generative AI (GenAI) have transformed the domain of human-computer interaction (HCI) (Muller et al., 2022). Sora OpenAI (Sora) is a recently announced advanced GenAI tool capable of crafting realistic videos based on textual input (Sun et al., 2024; Wang et al., 2024), which has sparked considerable interest and speculation regarding its potential impact on societal dynamics, governance structures, and social norms. Sora is equipped with sophisticated text-to-video generative abilities and holds the potential to revolutionize industries (Guan et al., 2024), reshape social interactions, and redefine the boundaries of human-AI collaboration (Waisberg et al., 2024).

As society begins to integrate Sora and similar generative AI tools into various domains, such as healthcare, education, and art, the need to comprehend the intricate dynamics of public perceptions and the broader societal implications becomes increasingly urgent. Effectively navigating the complex intersection of technology and society requires a deep understanding of the ethical, social, and cultural dimensions that accompany AI integration. Addressing initial reactions and concerns may mitigate the potential risks and safeguard the values and interests of individuals and communities after Sora's release. In this study, we collected Sora-related discussions from X (formerly Twitter). In addition to discussions about Sora-generated videos, we also included discussions related to a comparison between Sora videos and Midjourney images, and discussions on ethics initiated by an artist whose work was copied by generative AI, in our analysis. We additionally focused on governance solutions to people's expressed concerns. These aspects can potentially help envision regulatory frameworks to control the implications of Sora and GenAI in general.

We aspired to answer the following research questions:

- **RQ1**. What are the primary reactions, opinions, and concerns people express regarding Sora's integration?

- **RQ2**. What are the perceived implications of Sora for governance structures and social norms?

We contribute to the emerging literature on Sora and text-to-video GenAI in two ways. First, we explored people's reactions to and concerns about Sora's integration and their potential solutions to prevent its negative impact. Second, we proposed policy recommendations to address some of the concerns expressed, namely, AI watermarking, leveraging existing laws, enhancing Sora's accessibility, and adopting an entrepreneurial approach to regulation.

Related work: benefits, challenges, and public perceptions of Sora

Sora is an AI tool for text-to-video generation by OpenAI, whose demos show surprisingly good quality (In, 2024). Sora has blurred the lines between tangible and virtual realms and even '*prompted viewers to ponder their very existence*' (Cavalcante, 2024). Li et al. (2024b) created a benchmark to assess the quality of AI-generated videos based on their adherence to real-world physics principles and showed Sora's advantage over other text-to-image generation tools. In recent literature, scholars have predominantly focused on addressing the challenges, benefits, and public perceptions of Sora.

Benefits

Sora is poised to disrupt multiple industries, including movies, education, gaming, healthcare, and robotics, with potential benefits for individuals (Liu et al., 2024). According to Lin (2024), Sora may serve to broaden access to mental healthcare services. In the realm of robotics, Sora's capability to simulate intricate world dynamics proves crucial for advancements in autonomous driving (Guan et al., 2024), such as scenario generation (Li et al., 2024a). Democratization of art creation is

a notable affordance of Sora – '*Everyone will be a filmmaker*' (Cowen, 2024), which holds promise for broadening participation in artistic endeavours. Researchers speculate that medical professionals can use Sora to create informative videos for patients and the public (Waisberg et al., 2024). Sora can also become a potentially powerful tool for education and libraries, offering opportunities for diverse learning modalities, creativity, and critical thinking (Adetayo et al., 2024).

Challenges

Sora may introduce ethical quandaries concerning privacy and security, copyright, disinformation, and truth integrity. AI's enhanced video generation capabilities may ease the creation of adversarial attacks (Hannon et al., 2024). Copyright infringement is a paramount concern brought by Sora – it might have leveraged copyrighted materials such as videos for training (Karaarslan & Aydın, 2024). Other concerns raised by the researchers included misuse of AI-generated content (e.g., deep fake videos) and environmental costs. Rodríguez (2024) discussed the likely 'ultimate solipsism' introduced by the rise of 'self-movies,' where protagonists are the creators themselves, leading to implications of quality degradation. Moreover, the duality of Sora was expressed by Cavalcante (2024), who argues that Sora simultaneously indicates democratizing creation, i.e., anyone can become a visual storyteller regardless of their technical or financial status, and ethical dilemmas such as disinformation that puts the integrity of truth at risk.

Public perceptions

Researchers started to investigate public perceptions of Sora before its release. In the work of Mogavi et al. (2024), both people's envisioned applications of Sora and expressed concerns about this emerging technology were captured by qualitatively analysing Reddit discussions. People predicted that Sora would democratize video marketing and gaming content production, enhance educational content by generating illustrative videos, and substantiate people's creativity and storytelling abilities. In the meantime, Sora threatens creative and artistic jobs and perpetuates social biases from training datasets. We extend this work by examining discussions on a different online platform, comparing discussions around Sora and Midjourney, and focusing on strategies to regulate governance challenges.

Methodology

Data collection

To address our research questions, we manually gathered 292 comments under four popular posts on X discussing public perceptions of different aspects of Sora. The first and second posts contained Sora videos released by OpenAI – the first post was by a prominent tech influencer, while the second post was by OpenAI itself, potentially capturing views from different stakeholders. The third post compared Sora videos and Midjourney pictures created with the same prompts, emphasizing visual similarity. By including discussion about and comparisons with Midjourney, a text-to-image AI tool, we might be able to uncover how text-to-video models like Sora presented unique challenges to AI regulation and society compared to text-to-image models. The fourth post was by an artist complaining about Midjourney's copying behaviour, highlighting regulatory challenges. The four posts were chosen in a way that represents different stakeholders (OpenAI, tech influencers, artists, and the general public) and compares between different GenAI models (text-to-video vs text-to-image models). These posts selected exhibited high levels of engagement, indicated by the large number of views and comments. Descriptions of the posts are presented in Table 1. The comments were put in a hierarchical structure to reflect interaction dynamics. Our dataset includes 93, 139, 39, and 21 comments, respectively, from these posts. We arrived at this truncated list from the original number of comments by excluding comments that included little information. For example, although there were nearly 10 thousand comments under Post 2 by OpenAI by early March when we collected the data, most of the comments were short and not informative (e.g., '*This is great*').

No.	Poster	Topic	Post Time	Number of Views	Number of Comments	Number of Retweets	Number of Likes	Number of Bookmarks
1	Tech influencer	Sora videos	Feb 15, 2024	9.4M	437	3.2K	19K	8K
2	OpenAI	Sora videos	Feb 15, 2024	94.3M	9.9K	76K	142K	40K
3	Tech influencer	Sora videos vs Midjourney images	Feb 16, 2024	11.2M	674	3.9K	31K	11K
4	Artist	Midjourney copying	Mar 8, 2024	2.7M	127	9.2K	63K	3.2K

Table 1. Metadata about four X posts

Data analysis

We employed a thematic analysis (Braun & Clarke, 2012) to analyse the comments. The thematic analysis involved identifying recurring themes, patterns, and categories within the comments, enabling us to gain insights into the predominant discussion topics. Through qualitative analysis, we aimed to uncover the diversity of opinions, concerns, and perspectives expressed by users regarding Sora (RQ1) and its implications for governance and society (RQ2). Emerging themes included people's assessment of Sora videos, Sora's impact on industry and profession, governance challenges, and potential solutions to the challenges. Under governance challenges, subthemes included OpenAI's for-profit nature, replacing humans, data copyright issues, and fabricated content. We regularly discussed to reach a consensus on the themes and organize the themes and subthemes into a hierarchical structure. Below, we use anonymized quotes to illustrate our findings. We retained grammatical errors and typos in the quotes but partially censored swear words with [***] and removed emojis.

Findings

We generated a word cloud based on our collected data using NVivo R.14.23.0, highlighting discussions on art/artists, laws, and content, as shown in Figure 1, with larger font sizes indicating more mentions in the dataset.



Figure 1. Word cloud based on total mentions

Varied assessment of video quality

People had varied assessments about the video quality Sora afforded. While some regarded Sora's video generation capability as revolutionary, others pointed out failures in simulating the physical world and its lack of creativity.

Many people were amazed by Sora's video generation quality: '*These videos look so real*,' '*I cannot believe we are here already*.' Some people particularly praised the physical features captured in videos, '*the physics etc are just too good*'. Sora was seen as a revolutionary technology, '*It has come to a point where it is hard to imagine what AI will be able to do next*'. Some even suggested AI may have had access to alternative realities,

> *what if the AI has already access to alternative realities and is capturing those footages out there and that's the reason it looks so realistic, and we cannot see the difference.*

To some, the video quality was not as impressive, '*I don' think it even considers physics in that sense, like if it were a game engine*'. Failures in capturing physics were pointed out, e.g., '*Dog skins and snow doesn't follow physics correctly*'. Many people found that the walking woman's legs switched sides several times in a video. Some people thought the Sora videos were merely a product of the huge computational power or datasets. Some people were sceptical and thought OpenAI might have used videos that matched the prompt, '*How do we know its not just giving out unaltered training data that matches the prompt? a dog in the snow video is not uncommon.*'

When comparing Sora and Midjourney, people found they converged to similar output, '*This just shows that most foundational models are going to converge to a similar output. This should give some AI hype pause*'. Proprietary data was seen as a key to training high-performance models, whereas all models currently are trained on similar public data, '*It's now down to who has the most robust proprietary data. The issue now is all these models are trained off the same data sets*'. With all the AI-generated content online as potential training data, the convergence would only get worse, '*Especially as the internet becomes filled with "noise" in the form of AI outputs, convergence seems inevitable*'.

Impact on industries and professions

Many expressed the potential for Sora to disrupt the movie and content creation industries. This quote, '*Sora is incredible and scary*', demonstrated ambivalence.

People agreed that the movie industry would be directly impacted by video generation AI like Sora, '*Hollywood is about to implode and go thermonuclear*,' '*RIP the entertainment industry. In a few months, people will make "Avatar" on their phones*'. One person even imagined the collaboration between Sora and Midjourney, '*Midjourney can generate the thumbnails while Sora the movie*'. Sora may democratise movie making, making everyone an artist, which has unpredictable implications. One person raved about this democratization process,

> *thank you, OpenAI has literally democratized cinema. Now, everyone can be an artist and create their own movies going forward. This will drastically reduce the exorbitant expenses of Hollywood, especially for low-quality content with high costs.*

However, the democratization of movie creation may also lead to low-quality content,

> *democratize: Discard all talent through stealing people's work then allow anyone who don't know a thing about the art to create their own slop. Wow I love this new democratized cinema world.*

This impact may extend to the rest of the media production industry, content creators, and artists. On the bright side, Sora may enable writers to do new forms of storytelling,

> *perhaps one day regular people, as authors, will be able to create feature-length films from (only) their writing. It would be a refreshing restart to finally wrestle the art form away from Hollywood and allow real stories to be told again.*

Sora may democratize content creation for average people,

> *awesome tool. Tools like this integrated with no-code will truly democratize what regular people can create and push out online. This stuff is amazing and that exponential growth this year will be insane.*

The other side of this coin, however, is that Sora may cause job losses for artists, '*I don't think y'all realize how many artists you're f[***] over right now*'; the rise of AI content farms,

> *i would like to miss it because it threatens every single creative profession. Music creation, art, story writing, screenwriting, photography, and film. Unless you are at the highest production level where protections exist, it'll be impossible for human creators to get traction,*

and the frustration of art students,

> *AI is making me feel I'm wasting my time studying arts. At this point I fear my future is gonna be affected by this medium, the fact that this fear is getting stronger with each new i see about AI "progressing" sickens me, and I'm sure I'm not the only one.*

Some people recognized the unfair game between GenAI and human artists, i.e., GenAI has been stealing artists' work as its training data, '*It hasn't democratised anything. It has stolen people's data without consent to profit of it. Everyone could already be an artist some are just too lazy to*'. When an artist shared her experience of being copied by Midjourney, another artist was frustrated and questioned why AI evangelists could not see this dark side of GenAI, '*Breaks my heart. I'm an amateur artist. I'm sure I'd be pissed if other people bastardised my professional efforts. I simply can't see why the AI evangelicals don't see it*'.

Some were optimistic despite negative impacts, thinking Sora would not replace artists and content creators in the short term. Sora was not regarded as creative, '*It's never replacing real artists. Machines only do loops so that means that they only repeat what has been made. Only Humans are truly creative*'. Sora did not recreate the human touch, '*Even as machinery comes to the scene, there's always gonna be a demand for man-made stuff. It can never truly recreate that human touch so integral to the craft*'.

Governance challenges

Governance challenges discussed included OpenAI's for-profit nature, the authenticity of content, taking people's jobs, training data, power consumption, and the copyright of AI-generated content.

For-profit OpenAI

People critiqued OpenAI's for-profit nature: '*What shocks me is not just the Quality but the level of Secrecy that OpenAI was able to maintain around this project is insanely impressive*'. Others thought OpenAI did not aim to research AI for the public good but only pursued profits:

> *ah yes for a company supposedly dedicated to researching AI and bringing it to the public I sure love the level of secrecy involved at the company, and it's not at all terrifying that a company with this level of secrecy wants 7 trillion dollars for further AI development.*

People thought OpenAI did not care about Sora ethical concerns:

> *they just don't care. And they don't care about the slew of fake-videos and propaganda and computer-generated porn of real people and all of the people they'll hurt that way. They don't care about anything as long as they get to line their pockets.*

What is real? Falsified evidence, pornography, and propaganda

Some people claimed that they would no longer '*trust for videos on the web*' since '*nothing is real with AI*'. It was hard to tell AI- vs human-generated videos, '*Now people are releasing real videos*

saying it's AI, no one knows the difference.' However, others argued that things were not real even before AI,

> *you don't know what's real now.... If you're taking anything you see in the mainstream media at face value, then idk what to tell you. 99% of it is spin, bias, or even flat out wrong. AI used for nefarious purposes will just be an extension of what's already happening.*

People also expressed that humans have been bad at distinguishing between fake and real all the time, '*I mean, most people fall for the worst faked videos all the time...so, nothing's really changed*'. They thought technologies were not to blame, '*Lighters allow anyone to set a fire anywhere. This dangerous tool should be strictly controlled, otherwise we will all die in fires. how right it sounds*'. Technological solutions were believed to come to the rescue, '*A guy said years ago on some podcast i watched, as good as technology is at making something fake there's as good technology identifying if somethings fake*'.

Some people raised concern about falsified videos as court evidence, '*this is all fun and games until you end up in court watching a 60 second video evidence of yourself committing a crime you've never done.*' Another person replied with actional steps to falsify such a video,

> *now you can Google a random person's name, get just a profile pic and in a few short sentences create a video of that person doing anything you want. You only need a starting photo, rather than a full video of someone actually committing the crime to add their face to.*

Some fear that innocent people may be put into jail, '*This tech will put people in prison with falsified evidence. Not there yet, but not far off at all*'.

Video generation was also seen to facilitate easy creation of porn, '*Porn is about to get crazy!*', '*Porn industry about to go spiral into alternate universes of weird.*' With simple prompts, video generation AI such as Sora will be able to generate pornography,

> *I'm obviously missing something. Please explain to me in simple words why all this "hyper-realistic" imaging is interesting for anyone except criminals and pornographers. Do you really think "a fat girl from Ohio with her father's video-camera" (Coppola) will create a true masterpiece?*

Implications for child pornography are concerning, '*you are facilitating child pornography and you know that (: my question is how you people are able to sleep at night knowing what you're doing*'.

Video generation AI may interfere with elections with propaganda videos, '*Fabricated political propaganda will be all over the internet*', '*Looks like the 2024 election is going to be fun*'. Another person disagreed and argued that AI is not the one to blame but the underlying vulnerable democratic system,

> *if generated 10 second videos endanger democracy even though it's public knowledge that such video generation is possible, then don't you think the system was brittle to begin? I don't think generative AI is to blame for that.*

Replacing humans and small businesses

People complained that instead of benefiting humans, AI is often used to replace poor people,

> *AI is a good representation of how humanity fails to use innovation to progress. We COULD automate menial jobs, implement UBI, and give humans the freedom to create and invent, but instead we replace The Poors™ with barely functioning machines and then leave human to starve.*

AI kept squeezing people into smaller boxes of job options,

> *each new automation technology squeezes humanity into a smaller box; and then you see new standard of "normal" people complain about increases in clinical diagnoses for types of people who no longer have a modern purpose.*

There was a heated debate between human autonomy and technological evolution. On the one hand, '*You're literally hurting jobs with this*'. On the other hand, '*OK let's stop technological evolution because...jobs*'. Optimists believed AI could not easily take jobs requiring creativity and human touch,

> *analytical/automated/repetitive tasks are likely to be replaced regardless. Things AI can't do: imagination, creativity, drawing on life experiences, intuition, emotional connection, manual dexterity, etc. Those jobs will receive more boost in the future.*

According to them, AI could free people from monotonous jobs, '*AI could be used to reduce the human necessity to work monotonous jobs and allow us to explore our creativity*'. People were more assured after noticing similar generated content by Sora and Midjourney with the same prompt – they believed that GenAI was not creative and would not replace human workers,

> *i know folks are worried, but AI is going to just fool some CEOs into laying off in the short term till they realize no one will buy what is essentially uninspired clip art. Going to suck until the CEOs rehire folks though.*

Small Businesses without comparable computational power will likely be killed by tech giants such as OpenAI, '*OpenAI just can't stop killing startups*'. '*Sam Altman has more founders blood on his hand than any other CEO in history of mankind . . .*'. '*How are we supposed to keep working on our little SAAS product after this? I'd be less anxious even if aliens landed on Earth*'.

Copyright of input and output data

Training data was a major concern for many. Some feared that YouTube or TikTok videos have been used for training, which has been written into the platforms' terms of service,

> *my conspiracy is I feel TikTok, etc has been built to feed these ai tools and there's language in the terms that allows them to own certain rights to videos uploaded.*

Overall, people thought OpenAI was not transparent about where they obtained the training data and suspected they had scraped images and videos from the internet without consent or '*used copyrighted material without paying for it*'. People may unintentionally give up their data rights by not reading the terms of service closely. When one person asked, '*who approved their content to be scraped and put into the training?*' Others replied, '*You did. When you accepted the terms of service,*' '*Read the entire terms of service.*'

The similar output by Midjourney and Sora indicates the importance of proprietary training data like YouTube or X, '*All things being equal, all models eventually get to parity without proprietary training data sets.*' Some thought Google had an advantage for the massive amount of data it owns, '*Idk how the legal stuff would work but google is sitting on a f[***] goldmine.*' However, others pointed out that training on social media data may lead to bias in AI, as it learned human bias,

> *using social media to train your model on is very dangerous. It has been tried in the past and the ai always was trained by users to become racist within hours.*

Further, training on artists' and content creators' work, or normal users' content, was seen as stealing, '*Stop stealing people's work,*' '*So which animators art did you steal for this?*'

People were also wondering who would own the copyright of AI-generated content. Some thought future copyright law would favour AI developers over AI users, '*In time, law will benefit AI program*

developers. They will get the copyright, not the "ai artists". Mark my words'. Others did not mind the fact that they may not be able to copyright AI-generated content since the sunk costs were low with minimum human efforts in the generation process, '*Does it matter when you can generate new content again and again without the huge sunk costs that the traditional content industries have?*'

Potential solutions

People favoured regulation and AI literacy education to hold Sora accountable and limit its negative impact.

Law-enforced labelling of AI content

Law-enforced labelling of AI-generated content and tools for identifying AI-generated content are deemed important regulatory approaches,

> *laws are needed immediately. Any AI-modified video or image should be required to include a disclaimer symbol, mark, or tag of some kind that makes it obvious—even in films, artwork, and written content. Extreme penalties are needed for any who break these rules. Everything released should have a clear source or otherwise be unshareable. If we aren't extremely careful, we soon won't be able to believe our own eyes.*

Labelling of AI-generated content can facilitate copyright protection in the meantime.

However, many people found legal requirements to label AI-generated content infeasible for adversarial manipulation by users, '*People would just remove the watermarks or crop the videos*'; politicians' poor tech literacy, '*It doesn't help that a huge percentage of our most powerful politicians were too befuddled by new technologies to figure out how to program their* VCRs *20 years ago*'; enforceability of laws, '*Agreed, but like most regulations requiring watermarks/notations/etc. I worry that, while it may discourage the majority of problems, laws won't be enough to completely combat this type of conduct*'; and the slow process of legislation,

> *even if they do care, laws always take ages to create and implement in any capacity on purpose. Unfortunately, it's far too slow for how fast things are moving in the world of tech and stuff.*

A few people thought guaranteeing safety through strict law indicated '*taking out all the fun*' and '*restricting freedom of speech*'. They believed such guardrails might drive Sora unusable and instead, only illegal content should be limited,

> *many of the limitations you have put in are making the product unusable for the majority of users. The only content that should be limited is illegal content as what is "misinformation", "hateful content", and "bias" is different based on human interpretation, time, and other factors.*

Educating the public

In addition to law enforcement, some argued education was a more sustainable approach to mitigating the negative impacts of emerging technologies, '*No. We don't need more laws. We need citizens who can think and reason. The problem of "fake news" (or spin) has been with us for centuries*'. Education is important partly because AI technologies will be used to nudge people to do bad things, '*We need to educate the public on AI and make it a priority because this technology is going to be used to manipulate and socially engineer people into doing very bad things.*' Source verification is one important topic for education, '*Needs to be verified by reliable source (?) or a live happening with large audience present*'. The moral education could happen organically as had happened with prior technological revolutions, '*Which has almost always been true with science and technology - the science always comes way before the ethics and, sometimes, morality can catch up*'.

Discussion

Mogavi et al. (2024) utilized social media (Reddit) data and qualitative analysis to understand social media perspectives regarding people's concerns about its integration. People's concerns included threats to creative jobs, bias, harm to art, and unpredictability in creative workflows. In our study, people expressed a unique set of concerns, including (1) OpenAI's for-profit nature, which made their claimed safety measures less credible, (2) the blurred boundary between real and fake content, which had practical implications such as falsified evidence in court, pornography, and political propaganda during elections, (3) replacing human workers and small businesses, (4) copyright of input and output data, and (5) power consumption and environmental impact. We further gauged people's perceptions of potential regulation and solutions to these challenges, namely, law-enforced labelling of AI-generated content and educating the public. Below, we delineate the hype around Sora and propose governance recommendations.

Toward gauging tech hype around Sora

Prior to the release and widespread use of a product, and its early adoption, its mention in both media and technical texts helps it get attention and investments. Such attention is a double-edged sword as it also leads to the build-up of expectations to an extent beyond what is possible to deliver, leading to damaged credibility and reputation (Brown, 2003). The word 'hype' has been used to represent this cycle, and this most famously includes the Gartner tech hype cycle methodology developed by Gartner Inc., a technology research and consulting firm, in 1995; Gartner uses the hype cycle in its consulting reports to understand the evolution of the hype around emerging technologies. Gartner's hype cycle has five stages: Innovation Trigger, Peak of Inflated Expectations, Trough of Disillusionment, Slope of Enlightenment, and Plateau of Productivity. At the innovation trigger stage, the technology is at the proof-of-concept while in the second stage, there are very few practical implementations despite expectations peaking. Thereafter, the third trough stage is a sharp decline of interest and many business failures. It is only after this that the technology picks up again more maturely and finally reaches mainstream adoption in the final stage (Steinert & Leifer, 2010). We can infer that early expectations being inflated beyond what could be reasonably delivered is the expected norm of evolution of new technologies. It is further understood that the time needed for the full hype cycle might vary – while '*normal technologies might take five to eight years, fast-track technologies might only take two to four years for maturity*' (Dedehayir & Steinert, 2016). The bell-shaped hype curve produced is due to human irrationality, mainly the following: attraction to novelty and love for sharing, social contagion, and heuristic attitude in decision-making (Fenn & Raskino, 2008). It is also possible that there may be multiple peaks and troughs of visibility and expectations rather than a single one and this may not conform to a hype cycle at all (Dedehayir & Steinert, 2016; Steinert & Leifer, 2010).

Our work on the public discourse around Sora is a suitable representation of tech hype at the inception stage. The actual product has not been publicly released, and only a teaser has been provided by OpenAI. People were nevertheless excited about this proof-of-concept technology, evidenced by heated discussions on Sora-related topics on social media. They had drawn on past experiences with AI to form perceptions and expectations about Sora. We find a directionality to the expectations around Sora, with most comments taking a clear positive or negative position, and expectations and perceptions were logical/fact-based or emotional, resonating with prior research on tech hype expectations (Shi & Herniman, 2023).

Our findings indicate a sizable conflict on similar issues and features between people. A fundamental one is that people had conflicting assessments of Sora's creativity and video-generation quality: while some thought Sora was a revolutionary technique for capturing physical characteristics, others identified imperfections in physical simulation and noticed similarities between Sora-generated videos and Midjourney-generated pictures. These conflicts were also present in people's views on the desirability of the technology, its impact on employment, AI law

enforcement, its impact on content democratization, the extent of AI regulation, and trust in OpenAI.

Sora could also be viewed as a continuation of previous successful generative AI models. An interesting parallel can be found with virtual influencers, which are artificially generated entities acting as social media influencers, often using hyper-realistic AI-generated imagery, but still human-curated: many of the thoughts evoked in this study included blurring of lines between reality and fiction, democratization using AI, and competition between artificially generated content and human-generated content, resonating with quotes of long-term followers of virtual influencers (Choudhry et al., 2022). This is indicative of something more meta: the public sentiment, perception, and expectations of AI technology may be influenced less by the specifics of that particular product and more by the general zeitgeist of AI. The conversations about Sora were less about the features of Sora itself but more about its potential impact on society, jobs, and freedom, implicating that such fundamental debates on AI have existed in the public sphere (Calhoun, 1993) for some time, suggesting that such public opinion could potentially be harnessed in creating fair legislation as has been envisaged previously (Leane, 2010), and resonating with a more flexible legislative approach (Wiener, 2004).

Governance recommendations

AI watermarking

Building on results regarding legal enforcement of labelling requirements, it is necessary to clarify what the mandate is as much as it is to articulate procedures, rights, and responsibilities regarding enforcement mechanisms. One major governance recommendation from this work is labelling AI-generated content; AI watermarking is one approach to labelling by verifying authenticity and preventing misinformation or mislabelling (Fu et al., 2024; Regazzoni et al., 2021). It is also critical to protecting intellectual property (Regazzoni et al., 2021). Such a labelling approach places the burden of compliance on the organizations and researchers that produce generative AI models, services, and products rather than on end-users with varying degrees of technical literacy. This centralized compliance mechanism easily facilitates enforcement, disambiguates provenance, and provides clarity to users of AI systems and downstream information seekers. It also makes copyright or restrictions on fair use transparent to everyone who encounters the material.

Reapplying existing laws

Results from this project also suggest leveraging existing law to apply it in new contexts, such as anti-discrimination law to address issues of bias, harassment, and equity, along with obscenity and child protection laws to protect minors with respect to pornography. Creative applications of broadly and flexibly written laws as pertain to similar harms or governance challenges irrespective of technical system, jurisdiction or context, or regulatory domain can help to address emergent harms, as identified in this research, as well as around emerging technologies more broadly. Such efforts have already been useful in protecting victims of revenge porn and deepfakes (Levendowski, 2013; Levendowski, 2023). As uses of generative AI expand, the scope and context of challenges increase, but many of the fundamental rights challenged in these scenarios are already protected and existing precedents should be applied and extended.

Accessibility

The application of accessibility standards to AI systems is also emphasized by the results of this study. Concerns expressed about potential guardrails infringing upon speech or creativity, as well as potentially making systems unusable, can be assuaged by reconciling them with accessibility interests, given decades of research demonstrating the positive impact of accessibility standards and testing on usability for all (Petrie & Kheir, 2007). In this sense, by making generative AI more accessible, a variety of concerns and the needs of many populations can be addressed via compliance with existing, or extended standards. Concerns about labor replacement and under-

or over-representation in data and use of these system, can also be addressed with expanded user populations and the inevitable user innovation that accompanies accessibility (Raasch et al., 2008).

An entrepreneurial approach to regulation

Literature from environmental technologies and their regulation indicates that innovation may be reduced by regulation, but regulation has a positive effect on public health (Zhou et al., 2021) and encourages socially positive innovation, as with privacy (Lev-Aretz & Strandburg, 2020). Tradeoffs between public benefit and innovation resulting from regulation can also be seen in regulation perceptions expressed by the two largely non-intersecting interdisciplinary silos of law and technology, and law and economics (Butenko & Larouche, 2015). Law and economics assume that innovation is good for welfare, while law and technology is more critical and views innovation as largely exogenous to the regulatory process. While law and technology literature wish that law intervenes early in emerging innovations, law and economics posits the very opposite; it is thus desirable to balance these opposing perspectives (Butenko & Larouche, 2015). Technology regulation is a highly complex undertaking, and it requires an entrepreneurial approach with flexibility and empirical testing, constituting 'regulatory property' to protect against market failures but guard against overprotection, fostering an ecology that allows diffusion of regulatory innovations, adapting the prevailing theory of regulatory politics, engaging of global economies, and the rewarding of policy innovation (Wiener, 2004). This entrepreneurial and nimble approach to regulation will resolve the trade-off between 'unethical AI' and 'unusable AI' in the regulation of generative AI models.

Conclusion

Through thematic analysis of social media discussions about Sora, one generative AI tool, we gauged people's reactions to and concerns about Sora and regulatory solutions. We synthesized governance recommendations from public perceptions, including using AI watermarking to verify authenticity and protect intellectual property, leveraging existing laws to combat pornography and bias, and addressing the tension between ethics and usability of AI systems.

There are several limitations of this research. First, we leveraged qualitative social media analysis to probe perceptions of Sora's societal impact and ethical concerns. Potential biases exist in the sample selection process, and the chosen posts and stakeholders may influence the governance suggestions and interpretations of public concerns. We do not aim to provide a generalizable account of the complicated GenAI landscape. Future research could consider a larger-scale quantitative analysis to capture the evolving sentiment and perceptions around AI hype, and people's contrasting views as evidenced by our data. Second, bias may exist in our collected data due to groupthink and narrative-setting effects of early comments. Future research could conduct individual interviews to mitigate such bias.

About the authors

Kyrie Zhixuan Zhou is a PhD Candidate in the School of Information Sciences, University of Illinois Urbana-Champaign. He received his bachelor's degree from Wuhan University. His research interests are in human-computer interaction, computer accessibility, and AI ethics. He can be contacted at zz78@illinois.edu

Abhinav Choudhry is a PhD Candidate in Information Sciences, University of Illinois Urbana-Champaign, Illinois, USA. He received his master's from Cornell University and his research interests relate to the deployment of AI and software applications to health, financial literacy, and planning, particularly for older adults. He can be contacted at ac62@illinois.edu

Ece Gumusel is a Ph.D. candidate in Information Science with a minor in Computer Science at the Luddy School of Informatics, Computing, and Engineering at Indiana University, Bloomington. Her doctoral research focuses on user privacy dynamics in conversational text-based AI chatbots through mixed-method approaches. Her research agenda encompasses areas such as usable privacy, privacy compliance, human-computer interaction, and social informatics. She received her LL.M. (Master of Laws) in Intellectual Property and Technology Law at the University of Illinois at Urbana-Champaign, and her LL.B. (Bachelor of Laws) at Başkent University. She can be contacted at egumusel@iu.edu

Madelyn Rose Sanfilippo is an Assistant Professor in the School of Information Sciences, University of Illinois, Urbana-Champaign USA. She received her Ph.D. from Indiana University, and her research focuses on governance in sociotechnical systems. She can be contacted at madelyns@illinois.edu

References

Adetayo, A. J., Enamudu, A. I., Lawal, F. M., & Odunewu, A. O. (2024). From text to video with AI: the rise and potential of Sora in education and libraries. *Library Hi Tech News*.

Braun, V., & Clarke, V. (2012). *Thematic analysis*. American Psychological Association.

Brown, N. (2003). Hope against hype-accountability in biopasts, presents and futures. *Science & Technology Studies*, *16*(2), 3-21.

Butenko, A., & Larouche, P. (2015). Regulation for innovativeness or regulation of innovation?. *Law, Innovation and Technology*, *7*(1), 52-82.

Calhoun, C. (Ed.). (1993). *Habermas and the public sphere*. MIT press.

Cavalcante, D. C. (2024). Sora: The Dawn of Digital Dreams.

Choudhry, A., Han, J., Xu, X., & Huang, Y. (2022). "I Felt a Little Crazy Following a 'Doll'": Investigating Real Influence of Virtual Influencers on Their Followers. *Proceedings of the ACM on human-computer interaction*, *6*(GROUP), 1-28.

Cowen, T. (2024). What will the main commercial uses be for Sora?.

Dedehayir, O., & Steinert, M. (2016). The hype cycle model: A review and future directions. *Technological Forecasting and Social Change*, *108*, 28-41.

Fenn, J., & Raskino, M. (2008). *Mastering the hype cycle: how to choose the right innovation at the right time*. Harvard Business Press.

Fu, Y., Xiong, D., & Dong, Y. (2024, March). Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 16, pp. 18003-18011).

Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Li, Y., ... & Xu, C. (2024). World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*.

Hannon, B., Kumar, Y., Gayle, D., Li, J. J., & Morreale, P. (2024). Robust Testing of AI Language Model Resiliency with Novel Adversarial Prompts. *Electronics*, *13*(5), 842.

In, A. I. (2024). SORA: Unbelieve New Text to Video AI Model By OpenAI-37 Demo Videos-Still Can't Believe Real. *Sign*, *3*(01).

Karaarslan, E., & Aydın, Ö. (2024). Generate Impressive Videos with Text Instructions: A Review of OpenAI Sora, Stable Diffusion, Lumiere and Comparable Models. *Stable Diffusion, Lumiere and Comparable Models (February 19, 2024)*.

Leane, G. W. (2010). Deliberative Democracy and the Internet: New Possibilities for Legitimising Law Through Public Discourse?. *Canadian Journal of Law & Jurisprudence*, 23(2), 373-401.

Lev-Aretz, Y., & Strandburg, K. J. (2020). Privacy regulation and innovation policy. *Yale JL & Tech.*, *22*, 256.

Levendowski, A. (2013). Using copyright to combat revenge porn. *NYU J. Intell. Prop. & Ent. L.*, 3, 422.

Levendowski, A. (2023). Defragging Feminist Cyberlaw. *Berkeley Tech. LJ*, 38, 797.

Li, X., Miao, Q., Li, L., Hou, Y., Ni, Q., Fan, L., ... & Wang, F. Y. (2024). Sora for scenarios engineering of intelligent vehicles: V&V, C&C, and beyonds. *IEEE Transactions on Intelligent Vehicles*.

Li, X., Zhou, D., Zhang, C., Wei, S., Hou, Q., & Cheng, M. M. (2024). Sora generates videos with stunning geometrical consistency. *arXiv preprint arXiv*:2402.17403.

Lin, B. (2024). The Machine Can't Replace the Human Heart. *arXiv preprint arXiv*:2402.18826.

Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., ... & Sun, L. (2024). Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv*:2402.17177.

Mogavi, R.H., Wang, D., Tu, J., Hadan, H., Sgandurra, S.A., Hui, P., Nacke, L.E.: Sora openai's prelude: Social media perspectives on sora openai and the future of ai video generation. arXiv preprint arXiv:2403.14665 (2024)

Muller, M., Chilton, L. B., Kantosalo, A., Martin, C. P., & Walsh, G. (2022, April). GenAICHI: generative AI and HCI. In *CHI conference on human factors in computing systems extended abstracts* (pp. 1-7).

Petrie, H., & Kheir, O. (2007, April). The relationship between accessibility and usability of websites. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 397-406).

Raasch, C., Herstatt, C., & Lock, P. (2008). The dynamics of user innovation: Drivers and impediments of innovation activities. *International Journal of Innovation Management*, 12(03), 377-398.

Regazzoni, F., Palmieri, P., Smailbegovic, F., Cammarota, R., & Polian, I. (2021). Protecting artificial intelligence IPs: a survey of watermarking and fingerprinting for machine learning. *CAAI Transactions on Intelligence Technology*, 6(2), 180-191.

Rodríguez, E. Q. (2024). Toward Solipsism: The Emergence of Sora and Other Video Generation AIs in Audiovisual Creation.

Shi, Y., & Herniman, J. (2023). The role of expectation in innovation evolution: Exploring hype cycles. *Technovation*, *119*, 102459.

Steinert, M., & Leifer, L. (2010, July). Scrutinizing Gartner's hype cycle approach. In *Picmet 2010 technology management for global economic growth* (pp. 1-13). IEEE.

Sun, R., Zhang, Y., Shah, T., Sun, J., Zhang, S., Li, W., ... & Ranjan, R. (2024). From Sora What We Can See: A Survey of Text-to-Video Generation. *arXiv preprint arXiv*:2405.10674.

Waisberg, E., Ong, J., Masalkhi, M., & Lee, A. G. (2024). OpenAI's Sora in medicine: revolutionary advances in generative artificial intelligence for healthcare. *Irish Journal of Medical Science*.

Wang, F. Y., Miao, Q., Li, L., Ni, Q., Li, X., Li, J., ... & Han, Q. L. (2024). When does Sora show: The beginning of TAO to imaginative intelligence and scenarios engineering. *IEEE/CAA Journal of Automatica Sinica*, *11*(4), 809-815.

Wiener, J. B. (2004). The regulation of technology, and the technology of regulation. *Technology in Society*, *26*(2-3), 483-500.

Zhou, G., Liu, W., Wang, T., Luo, W., & Zhang, L. (2021). Be regulated before be innovative? How environmental regulation makes enterprises technological innovation do better for public health. *Journal of Cleaner Production*, 303, 126965.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Investigating privacy risks in open government data: an exploratory case study

Daniel Carter and Caroline Stratton

DOI: https://doi.org/10.47989/ir30iConf47287

Abstract

Introduction. As governments increasingly release large datasets to the public, concerns have arisen regarding the potential risks associated with the inadvertent disclosure of personal information. However, there are few empirical case studies of this phenomena.

Method. Taking a red teaming approach, we collected public, nonprofit reporting documents from the IRS, as well as a list of professional athletes. Using standard data processing techniques, we attempt to identify and evaluate the personal information that can be attached to these notable individuals in order to propose a framework describing how features of information, people and processes influence the potential release of personal information.

Findings. We find that name commonness plays a large role in determining the feasibility of matching records. We also find that personal information, including addresses and business relationships, are accessible in the dataset.

Conclusion. Our exploratory case study suggests several features of information, people and processes that influence the release of private information held in open government data. While limited, it suggests the value of further work in the area.

Introduction

In recent years, the open government data (OGD) movement has gained significant traction, promoting transparency, accountability, participation and trust (Janssen et al., 2012). However, as governments increasingly release large datasets to the public, concerns have arisen regarding the potential risks associated with the inadvertent disclosure of personal information(Graham et al., 2016). This paper extends the literature on open government data by presenting an exploratory study that evaluates the risk related to a specific case of government data and proposes an initial framework that can guide future work.

Specifically, we focus here on the risk that personal information might be inadvertently released in nonprofit reporting documents such as the IRS form 990. The risk scenario that we attempt to better understand is one in which a dataset collected from such a source might be used to identify personal information such as addresses and contact information for notable individuals such as celebrities (in this case, NBA players) or a list of less notable people that might be relevant in a specific context (e.g., journalists, activists or teachers). While such contact information is available through other public sources, such as property records, we choose to work with nonprofit reporting documents because they have a national scope and, for risk scenarios that rely on targeting individuals with wealth, likely represent a higher density of targets.

The generalized information context that we attempt to understand, then, is a large dataset that is not valuable as a whole but is rather assumed to contain some valuable information that must be located and extracted. In conducting a case study to better understand this phenomenon, we assume that an adversary would approach this as an information retrieval task and model our work accordingly.

Based on our findings, we propose an initial framework that identifies key attributes of information, people, and processes that may impact the likelihood of harmful data release as a result of OGD initiatives. By examining these factors, we aim to contribute to the ongoing dialogue about balancing the benefits of open government data with the need to protect individual privacy and prevent potential misuse of personal information.

Literature review

The literature on OGD has raised the possibility that privacy is a salient concern for governments and may be in tension with the aim of transparency (Graham et al., 2016; Janssen & van den Hoven, 2015). In general, while OGD conforms to the conditions that Janssen et al. (2012, p. 258) stipulate as '*non-privacy-restricted and non-confidential data which is produced with public money and is made available without any restrictions on its usage or distribution*,' the removal of all personally identifiable information (PII) from OGD, both that of residents who interact with governments and that of government workers, is far from uniform. In some cases, the value of an open dataset lies in its identification of people and their associated characteristics. The three most popular datasets on New York City's open data portal as of September 2024, for example, list individuals' names and their vehicle identification numbers (For Hire Vehicles dataset), individuals' names and scores for a civil service exam (Civil Service List dataset), and individuals' names and addresses associated with building permits (DOB Job Application Filings dataset) (City of New York, n.d.).

While scholars have speculated about the relationship between privacy and OGD, we find little identification or exploration of specific cases in the literature. For most OGD research, the concept of individual privacy is left at the hypothetical level.

In the years since OGD has grown in popularity, other phenomena that jeopardize individual privacy, such as the growth of data broker companies (Crain, 2018) and the prevalence of large-scale corporate data breaches (e.g., Newman, 2024; Srinivasan, 2017), have perhaps overshadowed interest in risk and OGD. While the impacts of selling and stealing data are certainly grave, the risk

posed by PII in OGD is one of hiding in plain sight, with potentially valuable personal information available to anyone capable of finding it. An older example illustrating potential harms of PII in OGD comes from Eightmaps, a visualization that showed donors to a 2008 California ballot measure to stop single-sex couples from marrying (Stone, 2009). De-identifying individual data is among the strategies that governments might take to reduce individual risk to privacy. Techniques for re-identification, though, may be powerful enough to nullify such efforts. Liu et al. (2021), for example, raised the possibility of re-identification for a dataset of car trips from highway toll-collecting, with license plate numbers removed. Because the dataset also contained information about location of departure, destination, and travel times, the authors argued that re-identification might be possible for some entries.

The potential for exploitation of PII found in OGD would seem to be against the ethos of the OGD movement, which has emphasized the transparency benefits and potential gains in efficiency for government and other stakeholders; however, one of the most important proclamations regarding OGD, Barack Obama's 2013 Executive Order (2013), declares that these data can and should lead to economic benefit, because enterprising people can '*develop a vast range of useful new products and businesses using these public information resources*.' A tension emerges here, then, between the commercial value of OGD and the potential for that value to be derived through the exploitation of PII found in government information. In this short paper, we explore risk around PII in OGD through a specific information retrieval task, building toward future work by asking:

What are the salient features of public information, people and processes that put personal information at risk?

Methods

To explore our research question, we adopted a red teaming approach (Mansfield-Devine, 2018) designed to understand how the release of personal information might be perpetrated by someone with a financial or personal incentive. We began with a dataset composed of Form 990-N submissions made publicly available from the IRS. Form 990-N is an abbreviated version of the standard form 990 which is used by nonprofit organizations with gross receipts of less than $50,000 to make annual reports. The abbreviated form 990-N data is made available in a single, delimited text file, as opposed to the full form 990 data, which would require more sophisticated search and join procedures. While we assume that there are differences in these datasets related to the information made available as well as other features such as the prevalence of professional document preparers, we chose the abbreviated set in order to facilitate an initial, exploratory study and because the ease of acquiring and working with the data might more closely align with the workflows of adversaries.

Our *form 990-N dataset* consists of 1.39 million records, representing annual reports from 2007 through 2023. The relevant information included in the data includes the principal officer's name and address as well as the address listed for the organization. We note that the full form 990 includes additional information that might be relevant to the topic, such as information related to revenue and board member compensation.

In order to explore the presence of information related to notable individuals in the dataset, we use a list of NBA player names obtained from basketball-reference.com, which includes all players who have ever played in the league. The dataset, which we refer to here as the *target dataset*, includes 5,738 players. Because the name of the nonprofit's principal officer is the only feature that can be matched, we do not include other features, such as date of birth, in our *target dataset*.

In order to explore the extent to which notable individuals can be identified in our *form 990-N dataset*, we first reduce the dataset to include only records that match a name in the *target dataset*. The resulting *matched dataset* includes 5,769 form 990-N records that might correspond to a

notable individual. However, some of these records will obviously represent false positives, or organizations with a principal officer who shares a name with a notable individual.

Because we are interested in features of the information as well as of the processes used to identify relevant personal information within it, we employ an evaluation process that balances thoroughness with feasibility, as follows:

1. **Ranking by Commonness:** We first ranked all the names in our dataset based on their commonness, from least common to most common using Hadley Wickham's babynames dataset (2022), which includes all first names used at least five times in the US, as provided by the Social Security Administration. This ranking allowed us to prioritize the evaluation of names that were more likely to correspond to notable individuals.

2. **Cutoff Determination:** The information contained in the public forms collected is insufficient to automatically verify if information is relevant (i.e., it corresponds to a notable individual); therefore, a more time-consuming, manual process must be employed. Because we assume that the value of information about players is randomly distributed (i.e., there is as much likelihood of value in the information regarding a notable individual that is easy to identify as one that is difficult to identify), we also assume that an efficient search strategy, such as would be adopted by an adversary, would focus efforts on data that is easier to verify and has a greater chance of corresponding to a notable individual. We discuss findings related to determining such a search cutoff below.

3. **Analysis:** For each record we evaluated, we determined the following:

    1. Relevance: Using information available online, are we able to determine if the named principal officer is a notable individual from our *target dataset*? We use a conservative approach and only record a match if we are able to locate explicit confirming information.

    2. Personal information: We evaluate if the information from the form 990-N record contains personal details such as a non-commercial address.

    3. Evidence of professional preparation: We evaluate if there is evidence that a professional entity such as a law firm or business manager prepared the form.

    4. Additional notes: We make note of any aspects of the form information that does not correspond to our analysis plan or warrants further attention.

Findings

We observe that both the *target dataset* and the *form 990-N dataset* exhibit an exponential distribution of name commonness (figures 1 and 2). We also observe that the *merged dataset* exhibits a markedly different distribution, with a greater proportion of results found to have higher name commonness (figure 3). Calculating the ratio of the proportion of values from the *merged dataset* to the *form 990-N dataset*, we find a roughly linear increasing relationship (figure 4), suggesting that the number of results that do not correspond to an individual in the *target dataset* will increase with name commonness. While this is not a surprising result, it has notable consequences for the general search problem described, as it suggests that an adversary would likely focus efforts on less common names.

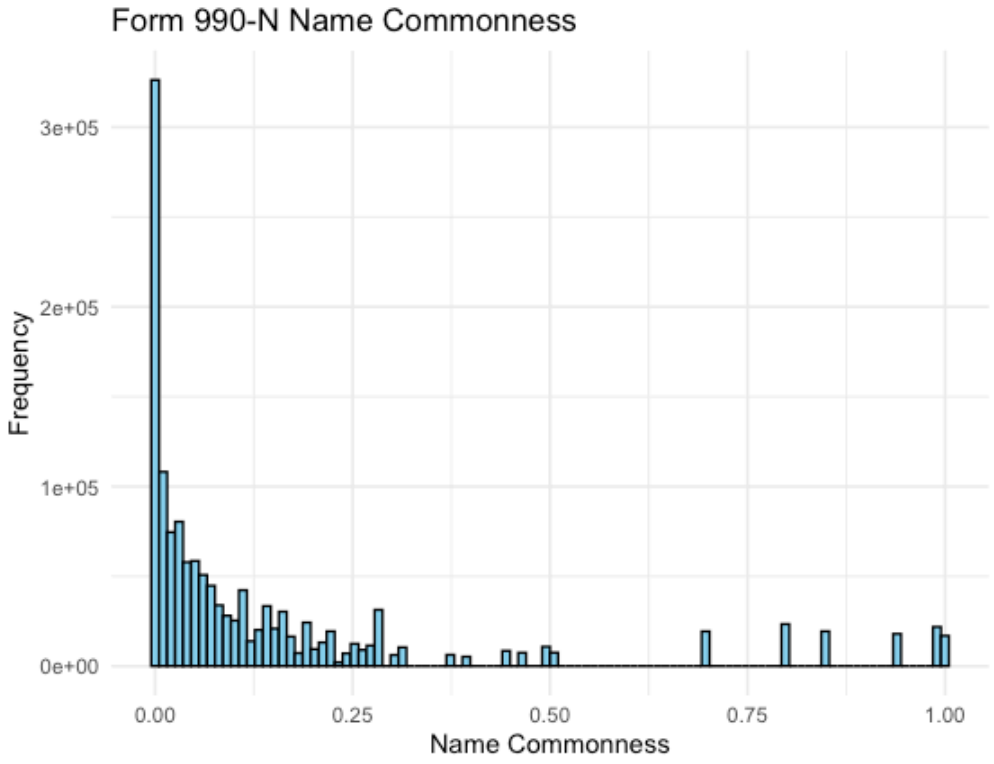


Figure 1. Distribution of name commonness in the Form 990-N dataset.

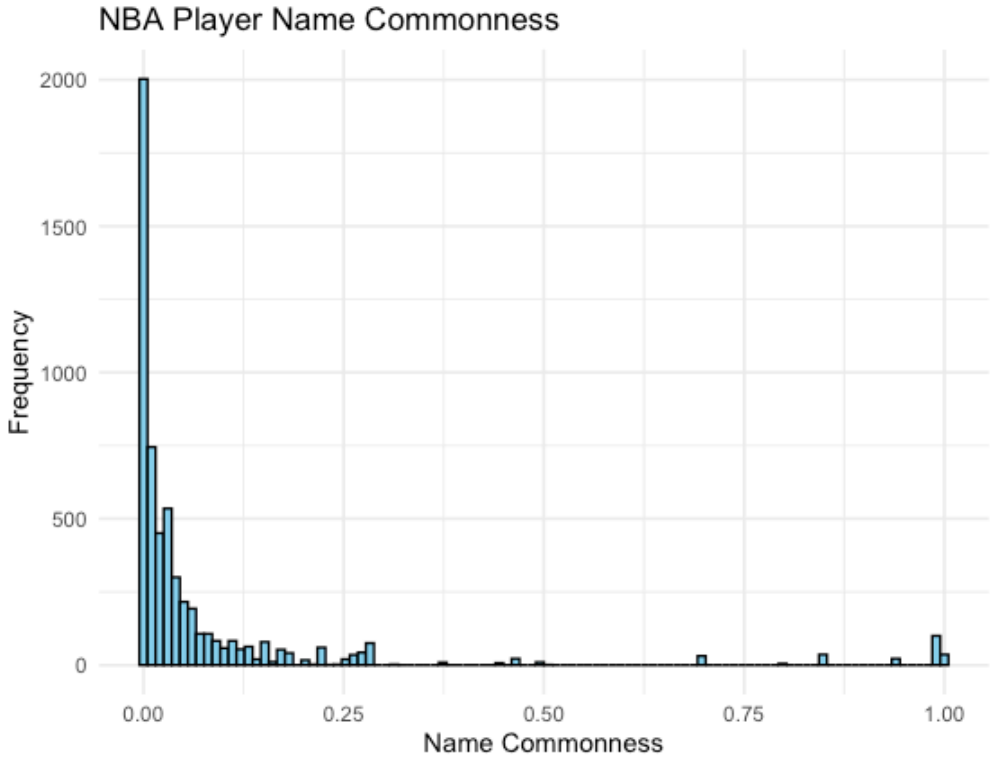


Figure 2. Distribution of name commonness in the target dataset.

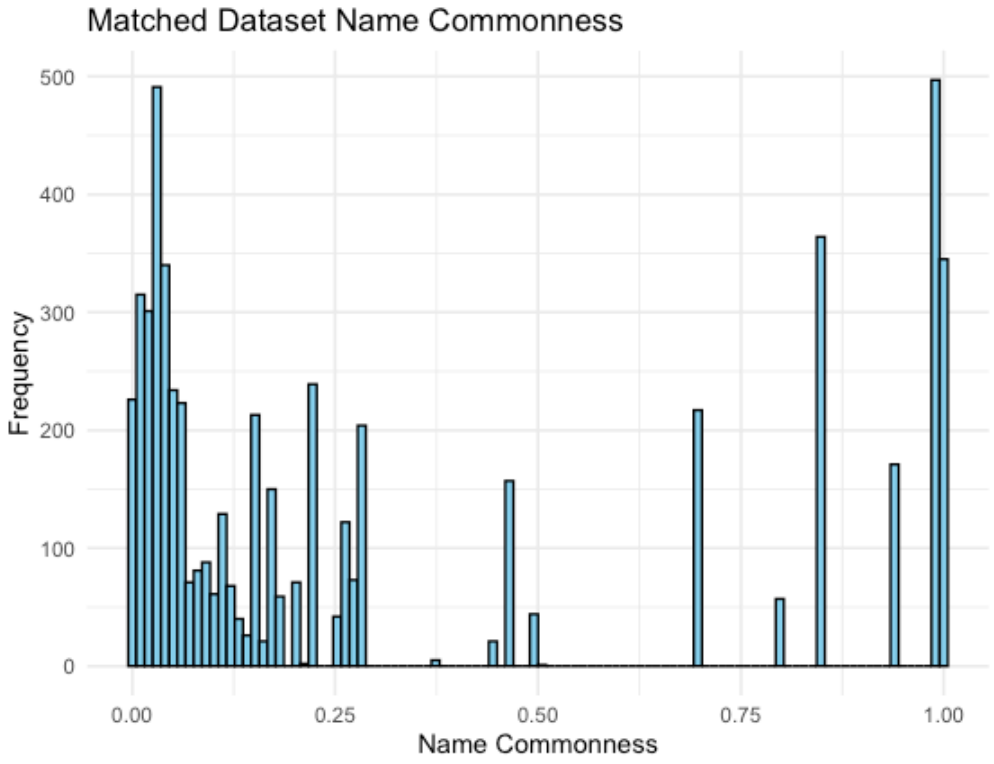


Figure 3. Distribution of name commonness in the matched dataset.

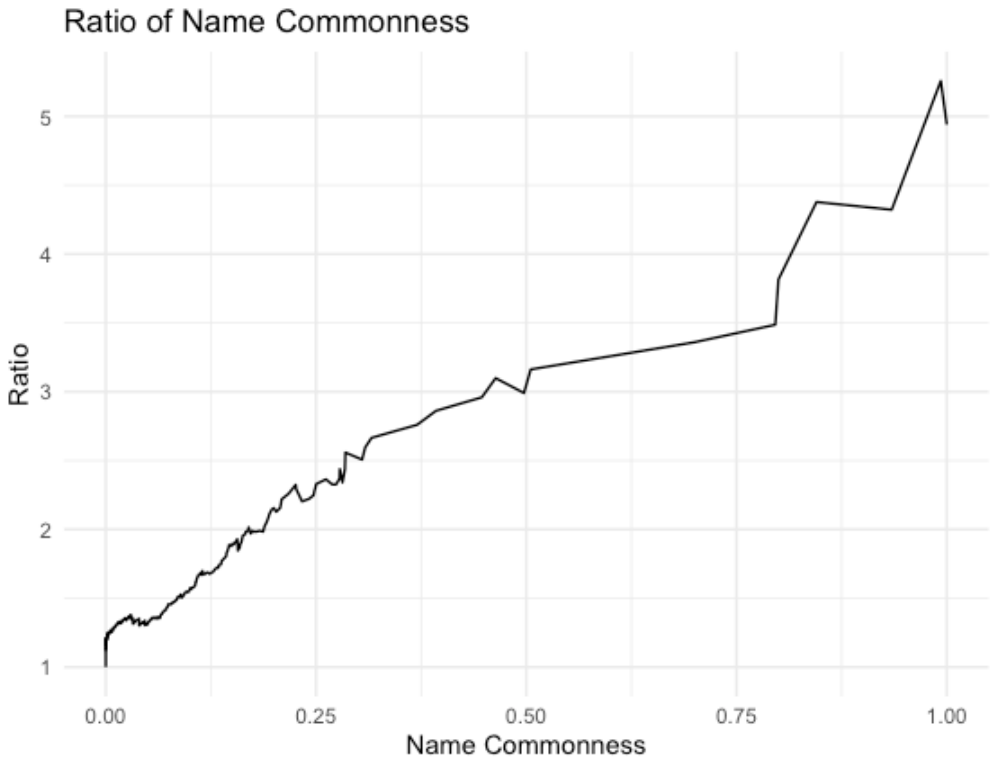


Figure 4. Ratio of the proportion of records, by name commonness, in the merged dataset and the Form 990-N dataset.

Because the purpose of this study is not to locate as much personal information as possible but to identify features of the information, people and processes involved that influence the risk of personal information release, we do not attempt to determine a precise cutoff (based, for example, in the ratio presented in figure 4) — indeed, given the highly contextual nature of the search task, determining such a limit is not realistic in this context. Instead, we evaluated records until we reached a point at which more than 100 consecutive records could not be matched to an individual from the *target dataset*, resulting in the inspection of 384 records. Figure 5 depicts the observed decrease in the rate of matches.

Of the evaluated records, we found that 13% matched an individual in the *target dataset*, with the rate of matches decreasing, as expected, as names became more common. Notably, we find that this rate decreases sharply, with just over 5% of records evaluated before matches became quite rare.

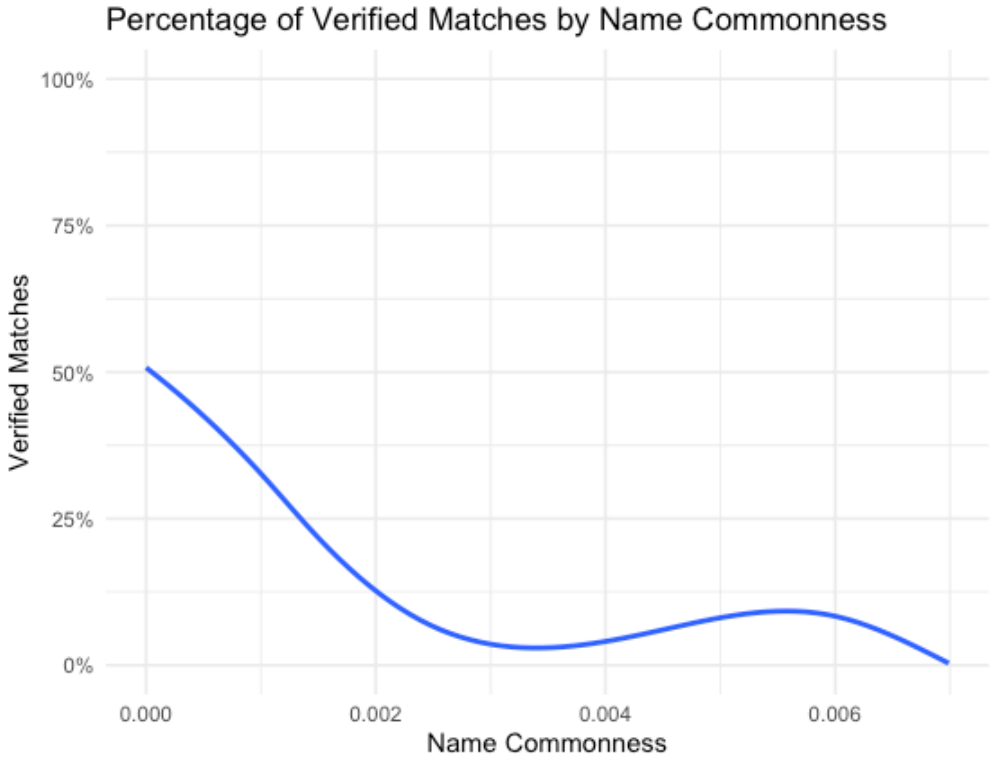


Figure 5. Percent of verified matched by name commonness

Of the verified matches, 52% included a non-commercial address, which we assume is likely the individual's residence. Notably, this information can be linked to other personal information, such as the value of an individual's home, which, while public, can be difficult to locate because property records are held at the county level.

36% of the matched records show evidence of the use of a professional service to submit the information. Observed professional services included both law firms and agencies that represent athletes. In all such cases, a commercial address was present instead of a residential address.

Discussion

Based on the results of our exploratory case study, we propose the following framework, which highlights features of the documents, people and processes involved in open government data.

Information characteristics

At a basic level, we observe that the fields included in the released public data play a role in determining the risk of personal information disclosure. While it's obvious that the inclusion of details such as social security numbers or health information would increase the risk related to disclosure, we also note that the absence of information that is commonly used to link records (e.g., date of birth) plays a marked role in reducing risk and makes the verification process more laborious.

Individual characteristics

We observe two features of individuals that likely impact the release of personal information through OGD processes: information literacy and name commonness.

Regarding information literacy, we observed several records in which fields were filled with information other than what was requested. This was most frequent with the form's Website field, which was sometimes used by individuals to enter an email address. In these cases, a lack of information literacy results in the release of personal information and provides another opportunity to link the record to additional information. The inclusion of personal addresses might be considered another example of a lack of information literacy. As we noted, professional preparation services, such as law firms, always use a commercial address, likely signaling that these entities believe there is an advantage to obscuring personal information. The inclusion of such information, then, signals an individual's lack of information literacy and understanding of how submitted information is, or is not, made public.

We also observe that less common names have a notable effect on the difficulty of verifying information in the specific search context described. This is especially obvious in our *target dataset* of NBA players, as a substantial number of foreign players come to the United States for this reason. As we discuss above, higher name commonality quickly makes verifying information impractical without automatic processing, making individuals with less common or foreign names much more vulnerable.

Process characteristics

Related to the process through which information is submitted, we observe that individuals who use a business intermediary such as an agent or lawyer appear much less likely to include personal information such as a home address. Instead, the address of the business appears in the public information; while this masks some of the individual's information, it can reveal other information such as business relationships that might be valuable in themselves.

Conclusion

This paper presents a needed empirical study of the risks associated with OGD. Taking a red teaming approach, we sought to better understand the features of information, people and processes that influence the risk of personal information disclosure. While we note challenges to the information retrieval task described above, notably in relation to name commonness and linking data, we also note that, with fairly little effort and using only public data, we were able to obtain the residential addresses of 50 NBA players, as well as other, non-public information such as the agency they contract with. The value of this information is highly context-dependent, but our study does suggest that a process similar to ours could be profitable to an adversary in some contexts.

Because this is an exploratory study, it has considerable limitations.

At a very broad level, we are constrained by the general difficulty of studying behavior (the exploitation of public information) that would likely be considered unethical and therefore difficult to access through conventional qualitative methods (see, for example, Carter et al., 2021). Our

approach here, red teaming, is adopted from computer science contexts in which the goal is not to describe a process occurring in the world but to imagine and circumvent such a process. While it is beyond the scope of this short case study to further explore the methodological and epistemological implications of adopting red teaming in the context of social science research, we believe that it has potential value, especially in the early, exploratory stages of research when there is a need for the creation of initial frameworks that can be validated and modified by further work.

While we believe that the data and processes, we adopt have features that likely generalize to other contexts, a small, single case is necessarily limited. Studies that, for example, survey public datasets and classify the kinds of personal information found in them could help guide such work. Additional case studies that purposefully address a wide range of contexts could validate and extend the framework presented above.

About the authors

Daniel Carter is an Associate Professor in the school of journalism and mass communication at Texas State University. They can be contacted at dcarter@txstate.edu.

Caroline Stratton is a social scientist who studies technology, society, and policy topics.

References

Carter, D., Acker, A., & Sholler, D. (2021). Investigative approaches to researching information technology companies. Journal of the Association for Information Science and Technology, 72(6), 655–666. https://doi.org/10.1002/asi.24446

City of New York. (n.d.). NYC Open Data. NYC Open Data. Retrieved September 13, 2024, from http://nycod-wpengine.com/

Crain, M. (2018). The limits of transparency: Data brokers and commodification. New Media & Society, 20(1), 88–104. https://doi.org/10.1177/1461444816657096

Graham, F. S., Gooden, S. T., & Martin, K. J. (2016). Navigating the Transparency–Privacy Paradox in Public Sector Data Sharing. The American Review of Public Administration, 46(5), 569–591. https://doi.org/10.1177/0275074014561116

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management, 29(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Janssen, M., & van den Hoven, J. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? Government Information Quarterly, 32(4), 363–368. https://doi.org/10.1016/j.giq.2015.11.007

Liu, C.-Y., Li, W.-P., & Tu, Y.-P. (2021). Privacy Perils of Open Data and Data Sharing: A Case Study of Taiwan's Open Data Policy and Practices. Washington International Law Journal, 30(3). https://digitalcommons.law.uw.edu/wilj/vol30/iss3/8

Mansfield-Devine, S. (2018). The best form of defence–the benefits of red teaming. Computer Fraud & Security, 2018(10), 8–12.

Newman, L. H. (2024, August 16). The Slow-Burn Nightmare of the National Public Data Breach. Wired. https://www.wired.com/story/national-public-data-breach-leak/

Obama, B. (2013, May 9). Executive Order—Making Open and Machine Readable the New Default for Government Information. The White House, Office of the Press Secretary.

Srinivasan, S. (2017, October). Data Breach at Equifax. Harvard Business School. https://www.hbs.edu/faculty/Pages/item.aspx?num=53509

Stone, B. (2009, February 7). Prop 8 Donor Web Site Shows Disclosure Law Is 2-Edged Sword. The New York Times. https://www.nytimes.com/2009/02/08/business/08stream.html

Wickham, H. (2022). Package 'babynames' [Computer software]. https://cran.r-project.org/web/packages/babynames/babynames.pdf

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Using the information inequity framework to study GenAI equity: analysis of educational perspectives

Sarah Zipf, Chuhao Wu, and Tiffany Petricini

DOI: https://doi.org/10.47989/ir30iConf47284

Abstract

Introduction. Generative AI presents opportunities and challenges for higher education, particularly concerning equity. Understanding stakeholders' perceptions of equity is crucial as AI increasingly influences teaching, learning, and administrative practices.

Method. The study was conducted in a large, research-intensive institution in the US. Participants (n=206) from diverse university roles responded to an open-ended question about how Generative AI affects educational equity. The responses were analyzed based on the information and equity dimensions (Lievrouw & Farb, 2003).

Analysis. Data were analyzed using a combination of deductive and inductive coding to identify key themes. The framework of information inequity underscores how disparities in access, skills, and ethical considerations create uneven opportunities for stakeholders to benefit from Generative AI, making these dimensions essential for understanding educational equity.

Results. Findings revealed differing focal points among the groups: faculty and staff concentrated on issues of physical and financial access to AI tools, while students placed greater emphasis on the ethical implications and value-based considerations of AI in education.

Conclusion. The study suggests that addressing AI equity in higher education requires a comprehensive approach that goes beyond improving access. AI literacy education should include skills development and address ethical considerations, ensuring that all stakeholders' concerns are met.

Introduction

Higher education struggles to keep pace with trends in technology, which has often trailed the industry significantly (Maslej et al., 2023). ChatGPT reached 100 million users in a mere matter of two months, which is exponentially less time than past services reached the same number of users, like Instagram and Facebook (Heath, 2023; Porter, 2023). Employers today want college graduates with experience and skills in GenAI tools (Cengage Group, 2023). Yet, the slow tendency for higher education to adapt to new technologies means a growing inequity is emerging between those who can use the tools effectively and those who cannot.

AI and equity have become a pressing concern, highlighted by organizations like the Bill and Melinda Gates Foundation, which considers AI equity among its top priorities. The foundation emphasizes not only equitable access to AI technologies but also the importance of co-creating tools that foster fairness and inclusion (Suzman, 2023). While the biased output of GenAI models is often cited as being an issue (D'Agostino, 2023), the issues of educational equity are much larger. While the last two years show an increase in research to understand the potential impact of AI technologies, particularly in the sphere of higher education (Petricini et al., 2023; Wu et al., 2024; Zipf et al., 2024), still missing from this growing body of literature are answers to the questions about equity (Bond et al., 2024).

The study of equity in the use of GenAI within higher education cannot be disentangled from the broader context of existing inequities that pervade society, higher education, and technology. These inequities manifest in multiple layers, influencing who has access to AI tools, who benefits from their usage, and whose voices are amplified or marginalized through these technologies. Central to these inequities is the concept of information inequity, which highlights the uneven distribution of access, skills, and agency in using AI tools, often reflecting, and reinforcing broader societal and educational disparities. The introductions of many past technologies have often led to both advocates and resistors, including the support and disdain for tools like the blackboard, calculators, and even the Internet, partly because the power dynamic between the instructor and student changed (see Davidson, 2022). As AI changes the power dynamics of knowledge, higher education structures will undergo fundamental and deeply transformative shifts that will lead to the 'privileging of certain forms of knowledge and communication' (Kurban & Şahin, 2024, p. 34) . Higher education has the potential to reduce or increase inequities of GenAI use for all faculty, students, and staff.

Artificial intelligence continues to grow in use and acceptance, making this a timely investigation for how key groups in the higher educational milieu perceive the impact AI has on equity. Our guiding research question was: What perceptions do students, faculty, and staff hold on the impact AI has on educational equity?

Examining equity and inequity in *GenAI usage and* higher education

Defining equity

Levinson et al. (2022) note a nearly universal commitment in higher education to equity initiatives while, at the same time, a clear definition of equity is absent. Often confused with equality, where everyone is treated the same, equity in education is about recognizing and addressing the specific needs of students to ensure fair access to opportunities. According to the American College Personnel Association, equity involves 'creating opportunities for equal access and success by addressing barriers and disadvantages that exist due to varying social identities and backgrounds' (Quaye et al., 2018, p. 4). This approach actively works to eliminate barriers and maintain an environment where all students, regardless of their differences, have the opportunity to succeed, grow, and achieve their individual goals.

Equity in education is a complex and dynamic concept, often influenced by political ideals. In one study, researchers investigated the term equity and concluded that while there is a growing body of work on equity, there is no established line of research or strong thematic connections, which indicates that this is a very diverse and fragmented field (Jurado de Los Santos et al., 2020). Equity in higher education is closely tied to access (see Wanti et al., 2022), which is often understood as the opportunity to attend college. However, this view of equity overlooks the complexity and nuances of what access truly entails. For example, Amaral argues that discussions of equity should encompass considerations of fairness and inclusion (Amaral, 2022). Embedded within the concept of equity are assumptions and questions about the roles and interactions of various stakeholders in a complex and multifaceted system.

AI's potential to facilitate equity

While AI has been criticized for perpetuating biases when poorly designed, it also has the power to foster equity in meaningful ways. AI has already shown promise in promoting equity in sectors such as law enforcement and the judicial system (Rigano, 2018), healthcare (Dankwa-Mullan, 2024; Schwarzman College of Computing, 2021). This technology carries the potential to address systemic inequities, including gender, race, and students with disabilities, found in educational systems.

The U.S. Department of Commerce recognizes AI's potential ability to enhance diversity, equity, inclusion, and accessibility (U.S. Department of Commerce, 2023). AI has the potential to improve job opportunities and learning experiences for individuals with disabilities. GenAI can help students find resources and draft accommodations (Lyerly, 2023). Other AI tools can assist students with low or no vision in using screen readers with a high level of detail (Brewer et al., 2020). Platforms like Knewton use AI to provide personalized learning experiences, potentially improving equity by ensuring that each student has access to the support that they need. When technology is introduced and controlled by certain groups, it creates 'normative implications that demand critical attention' (Grimes & Feenberg, 2015, p. 4). The on-going criticism of cheating and plagiarism around GenAI tool invalidates the potential for students with disabilities to have a more equitable education experience.

AI's adaptability makes it possible to tailor work environments and educational settings to the needs of diverse groups, further supporting inclusion and addressing racial inequalities. As Sahota (2024) writes in Forbes:

> *In a world grappling with racial inequalities, artificial intelligence* (AI) *offers promising tools for fostering equity and inclusion. By identifying and addressing biases in various areas, people are using* AI *to dismantle systemic barriers and promote racial equality.*

Together, these examples show that there is a wide range of potential applications of AI to promote equity across multiple areas of society. Whether by addressing gender biases, supporting individuals with disabilities, or promoting racial equality, AI offers promising tools for dismantling systemic barriers and creating more inclusive environments.

AI's potential to exacerbate inequities

Despite its potential to promote equity, AI also risks deepening inequities within educational systems. As D'Agostino (2023) highlights, one of the significant issues is the AI literacy divide. Some students are already familiar with and skilled at using AI tools, while others—often from underrepresented or low-income groups—may lack access to these technologies or the guidance needed to develop such skills, creating a disparity in who benefits from AI-powered learning tools. Some students can afford access to premium versions while others cannot (D'Agostino, 2023). The paid or premium versions of AI tools offer advanced features and more personalized learning

experiences. Relatedly, institutional-level access is also an issue, as some schools will be able to afford to provide access to faculty, staff, and students while others will not.

On a broader scale, Western higher education paradigms have traditionally emphasized information transfer—lectures, memorization, and exams—over creativity, critical thinking, and experiential learning. While AI holds the potential to transform this model, it also risks reinforcing the current power structures that privilege certain types of knowledge and learning, potentially sidelining marginalized voices. The deeper issue here lies in the historical and systemic inequities that already exist in society.

Education systems, despite being intended to mitigate these disparities (Amaral, 2022), often replicate them. AI technologies, influenced by societal biases and controlled by groups with significant power, can reflect these inequalities. Information inequity, or the uneven distribution of access to knowledge and tools, becomes a critical factor in this context (Lievrouw & Farb, 2003a). As AI becomes more prevalent in education, the gap between those with access to information and those without widens.

Technological equity concerns

The deployment of Generative AI (GenAI) tools in educational settings illustrates critical aspects of **information inequity**—the unequal access to, use of, and benefits from information technologies. Information inequity extends beyond physical access, encompassing disparities in digital literacy, socio-economic status, and the structural hierarchies embedded in technology deployment. For example, Educause reports that licensing agreements often privilege faculty and staff over students, with over half of institutions providing no AI literacy programs for students compared to 34% for faculty (Burns & Muscanell, 2024). This gap in institutional support reinforces inequities in who is prepared to use GenAI effectively.

Information inequity is further exacerbated by the AI literacy divide. Students with limited broadband access, those attending schools with AI bans, and those unable to afford premium AI tools face significant disadvantages compared to peers who can access and utilize these technologies. This divide creates a hierarchy of information access and use, limiting the ability of underprivileged groups to fully engage with the transformative potential of GenAI. These inequities reflect broader systemic issues, where educational access remains stratified by socio-economic and institutional barriers.

AI's reliance on biased datasets further compounds information inequity. By reproducing societal biases, GenAI systems risk marginalizing underrepresented voices and reinforcing existing power dynamics. For instance, algorithms in areas like language processing or decision-making can produce inequitable outcomes, disadvantaging groups already on the periphery of educational systems. These biases underscore the need to critically examine how information is curated, accessed, and applied through AI tools in educational contexts.

Methods

Important to this study is the definition of 'GenAI usage'. We define this term to mean interacting with tools to refine questions, seek answers or solutions, assist in brainstorming or idea generation, or any other action within a designated tool.

The site location for this study was a large, research-intensive, historically white institution (HWI), located in the mid-Atlantic region of the United States. Selected items are reported in this study as part of a larger research project. After approval from the institutional review board, we sent emails to unit leaders across the institution to share with faculty, students, and staff in their areas. The emails included a short introduction, a link to the Qualtrics survey tool, and implied consent and minimal risk with participation information.

Participants answered the open-ended question, 'How does generative AI impact educational equity?' Responses were analyzed with a combination of deductive and inductive thematic coding, in which the latent level was used to underscore and connect deeper meaning to the responses (Braun & Clarke, 2006). The codebook was developed a priori based on the five dimensions of information equity by Lievrouw and Farb (2003) (Table 1). We considered *Access* to mean the ability to pay for, use, or reach the GenAI technology and concepts related to socio-economic status and ability to obtain digital technology and devices. *Content* meant pertaining to the output, data, or deliverable from a GenAI tool, whereas *Context* was where GenAI tools were located within a learning environment and higher education, implying change to society, higher education, teaching and learning, and social systems. *Skills* related to the teachable abilities or knowledge about the use of GenAI tools and knowing how to perform actions and tasks. We also used *Skills* when people described using GenAI in different ways, either beneficial or harmful. Lastly, *Values* implied the principles or standards that govern personal and societal belief and behavior; determining if something is good or bad, beneficial, or harmful, right, or wrong, and highlights the ethics and morality of using the GenAI tool. By using these dimensions, we were able to maintain an equity lens on the responses.

Our data suggested additional dimensions not represented in Lievrouw and Farb (2003). For those responses unrelated to equity, we used N/A. Additionally, our data showed people were unsure of how equity matters or plays a role with GenAI tools and applied the code *Uncertainty* in those instances.

We utilized a thematic analysis approach to the data. Thematic analysis is a flexible analysis for any type of qualitative data that is not bound by a certain theory (Braun & Clarke, 2006). Data was loaded into an Excel spreadsheet and shared with the team. The research team discussed the five dimensions, and two researchers coded independently during round one, resulting in an unacceptable level of inter-rater reliability ($\kappa = .57$). The two researchers discussed the codes, agreeing on the meaning and extension of the dimensions, and re-coded the open-ended responses with an acceptable level of inter-rater reliability of ($\kappa = .89$).

Code	Description
Access	Ability to pay, use, or reach; related to socio-economic status, obtaining digital technology and devices
Content	Pertaining to the output, data, or deliverable from a GenAI tool
Context	Related to the location and idea of GenAI within a learning environment and higher education
Skills	Teachable abilities or knowledge about the use of GenAI tools; actions needed to perform tasks, such as prompt engineering
Values	Principles or standards that govern personal and societal belief and behavior; determine if something is good or bad, beneficial, or harmful, right, or wrong
Uncertainty*	Unsure of how equity matters/plays a role
N/A*	Equity is not mentioned in the response

Table 1. Codes adapted from Lievrouw and Farb (2003b) in alphabetical order (*Designates additional codes by researchers)

Limitations

Limitations of this study exist. Data that was coded by our method may not yield the same results if a different lens is applied. For example, some data were not returned as codable because the response had nothing to do with educational equity but rather served as a commentary on teaching and learning in general terms. While still very valuable to the larger discussion around GenAI, these points were not included in our analysis. Comparison to institutions with different faculty, student, and staff populations should be made with care. Our site location provides paid versions of GenAI tools to the entire community, which not all institutions can afford. Likewise, institutions with

different policies, either prohibiting or integrating GenAI tools in the classroom, institutional strategy, or operations, might find other concerns related to equity left unknown by our sample.

Results

Sample demographics

A total of 374 participants participated in the study, among which 206 provided valid responses to the open-ended question regarding equity. This paper focuses on 206 open-ended responses only. No duplicate responses were found in the data. The distributions of respondent demographic information are summarized in Figure 1. The majority of the respondents are university staff (48.5%), White (78.2%), and women (58.7%).

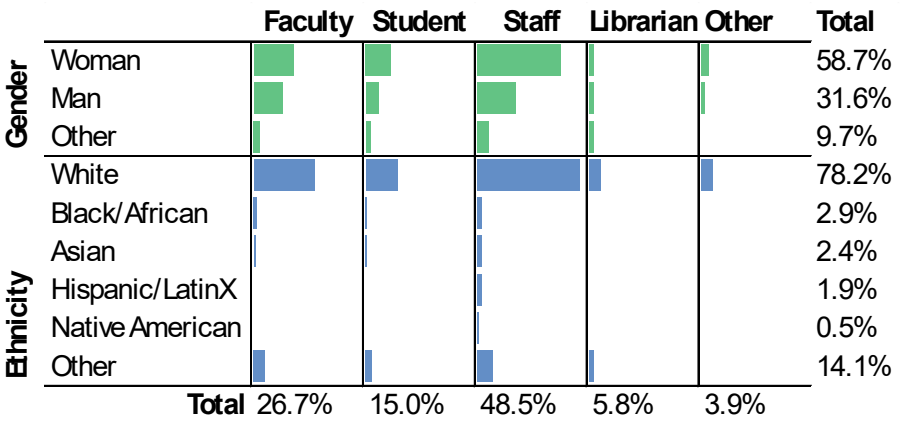


Figure 1. Participants demographic distributions

Codes

We received 206 open-ended responses. Staff had the highest representation (n=100), followed by faculty (n=55), students (n=31), librarians (n=12), and others (n=8) (i.e., postdocs).

Our data show differing opinions from students and staff about educational equity and the fact that multiple layers of GenAI and educational equity exist in higher education. *Access* was coded most frequently (n=42), followed by *Skills* (n=38), *Uncertainty* (n=36) and *Values* (n=33). *N/A* (n=16), *Content* (n=19), and *Context* (n=12) had fewer coding frequencies (Table 2). The codes show different foci by sub-sample; although *Access* was mostly coded for faculty, librarian, others, and staff, only one student mentioned this in their response to educational equity. On the other hand, 42% of the students who responded indicated *Values* in their responses, compared to zero in librarians. Faculty responses and staff mentioned *Skills*, and faculty, librarians, and others were *Uncertain* about how GenAI tools matter to educational equity.

Role	Access	Content	Context	N/A	Skills	Uncertainty	Values	Sub-Sample Total
Faculty	22% (12)	7% (4)	7% (4)	5% (3)	24% (13)	24% (13)	11% (6)	100% (55)
Librarian	33% (4)	17% (2)	0% (0)	8% (1)	8% (1)	33% (4)	0% (0)	100% (12)
Others	25% (2)	0% (0)	25% (2)	0% (0)	0% (0)	38% (3)	13% (1)	100% (8)
Staff	23% (23)	9% (9)	5% (5)	16% (16)	20% (20)	14% (14)	13% (13)	100% (100)
Student	3% (1)	13% (4)	3% (1)	19% (6)	13% (4)	6% (2)	42% (13)	100% (31)
Total Code Frequency	42	19	12	26	38	36	33	206

Table 2. Code Frequency by Respondent Role as Percentage of Sample Total

Based on the results from the coding, we found patterns in the data, and grouped the results into three themes (Braun & Clarke, 2006). First, we discuss the perspectives on values and ethics of GenAI use and the impact these concepts have on educational equity. Second, coded data shows people are unsure of how GenAI might influence educational equity, suggesting a surface level understanding of the technology. Third, we describe the forms of access shared in the open-ended responses. When combined, these findings suggest a more nuanced understanding of educational equity and GenAI is needed.

Perspectives on values and ethics

This section summarizes and compares participants' perspectives on values and ethics related to GenAI's impact on educational equity. Our data show different perceptions of how GenAI impacts educational equity based on the participant's specific role at the university. While students made up a small percentage of survey participants, they rarely responded to the question in any way related to access. Instead, students' responses often focused on *values* or the concepts of good or bad, right, or wrong, when applied to the effort of work. Students suggested using GenAI for studying is acceptable, 'but if it's used for cheating on assignments, then it should turn into a violation'. One student said the use of GenAI was 'unfair' because 'honest students' spend more time doing their work and another said 'students that work hard to create original work are at a disadvantage' compared to students who use GenAI. Some students considered how their instructors might use GenAI and felt the technology should not be used to 'complete an assignment or teach a class,' and others worried 'it may be used to create bland lessons' if instructors are not careful. One student linked the concept of values to societal beliefs, suggesting that while GenAI tools might offer valuable assistance to non-native English speakers, they could ultimately have 'an incredibly negative impact on educational equity' by encouraging complacency and creating an illusion of equity.

Staff did not share the same thoughts about *values*. Staff were concerned about how GenAI 'reflects the prejudices/biases of its human creators' or about the 'anti-Blackness being baked into the design' of predictive analytics. One staff member said educational equity would be impacted because of the 'biases built into AI… [and we're] rushing ahead, not considering what we're creating'. Other statements of values suggest that GenAI should be used 'correctly' or 'thoughtfully and intentionally' implying the opposites of incorrectly, superficially, or unintentionally will impact educational equity, though the outcomes are unclear. One staff member said, 'AI is the antithesis of education,' finding the question about educational equity 'to be contradictory'. Only two faculty responded that it was 'unfair' for students to use GenAI and implied that some students 'have too much personal integrity to use AI.' Lastly, one faculty mentioned that AI is 'unethical' even without

considering the 'environmental impacts in terms of energy consumption and greenhouse gas emissions.'

Related to values, ethics, and biased output appeared in *content* responses. *Content* or the output, data, or deliverable from the GenAI tools, was described as largely based on 'biased' data that 'has the potential to reproduce inequities that already exist'. One librarian called for a critical review of AI tools and stated:

> *Generative AI tools have unethical practices baked into their DNA [...] these tools are largely developed in an industry that is dominated by cisgender, heterosexual, White men -- and consequently the bias that this group can exhibit (overtly or unwittingly) becomes part of the machine learning/media generating process.*

As explained by one faculty member, when AI tools 'are trained on biased data then that is what they will generate (e.g., 'picture of college student' leads to pictures of white people),' which can be rectified with appropriate training to 'open pathways to diversity'. Several responses stated, 'garbage in, garbage out,' referring to the quality of output and the general concern for biased or offensive material.

One student wrote:

> *If I was in a class that was not focused on the usage of AI and learned that the teacher used AI for a chunk of their exams or lessons, I would be worried about the validity of my education.*

The implication of the student's concern is that faculty's use of AI in the classroom carries a different type of academic integrity component that will impact the value of their education. As seen in the ranked items, concerns of blame and trust in telling the truth are concerning. One student wrote that AI 'makes professors think that an essay would be AI generated and not believe the student when asked,' which is a sad indication of how students see AI impacting educational equity.

Multiple forms of access

This section explains the multifaceted concept of access and how it influences the equitable use of GenAI tools as reported by our participants. Staff's open-ended responses were the largest portion of our sample (n=100) and, like faculty and librarians, were concerned about *access* as it related to socio-economic status and the ability to pay for subscriptions. For instance: 'financially affluent students may be able to purchase the more advanced AI tools while less financially affluent will rely on the free version that aren't as robust'. The free version might be all that is needed, as one staff member wrote 'It's about access, as long as there are free tools there is no impact in my opinion'. Several staff respondents used the term 'have and have nots' implying a divide between those students who have the financial means to pay for the AI tools and those students who do not. One staff member thought of a larger impact and shared:

> *For AI apps or services that charge a fee to access or use, some students who come from low-income families may not be able to afford access while still in high school and into college, putting them at a disadvantage of how to use it, or to be able to gain any educational benefits from it -- putting them behind their peers who come from better financial circumstances and wealthier school districts.*

Thinking beyond the college campus, this faculty member uncovered a systemic issue of *access* to AI tools. The impact on educational equity was mostly positive, provided 'everyone has access' and 'there are free tools'. Overall, the financial challenges associated with the cost of these tools concerned faculty, librarians, staff, and students.

Access also meant obtaining or using technology, such as hardware and the AI software platform. Many responses about the impact to educational equity implied equality, or the same, access to the tools. One faculty member thought AI would be 'the same way that open access resources impact educational equity,' meaning equal or shared level of access to materials. In this manner, cost was not the only inequity, because 'students who actually understand the aspects of AI and could have [had] better access to technology would get more from the AI' than other students and 'all AI tools require access to the internet, and some people (staff, students, and faculty) have more reliable access to the internet and better personal devices (PCs, laptops, mobile phones, and tablets) than others'. One staff member linked this to the 'digital divide, the best tools cost money and not all students can afford these tools' and 'not all folks have access to a computer'. A student said that AI will not impact educational equity because 'given the availability of computers on campus as well as personal smart devices like phones and tablets,' people have the materials necessary to use the technology. Still related to financial stressors, *access* is differentiated by the materials needed to use the AI technology.

One staff member felt that AI tools would create employment inequity, sharing that 'fully remote employees can use the money they save from daily commutes to purchase an AI tool and further reduce their work time paid by the University' and obtain a second job. While outside the focus of educational equity, the concern shown by the staff person suggests staff may face different forms of inequity with the introduction of AI into their professional work.

Access can also mean usability for those with other abilities or disabilities. One faculty thought, 'if carefully used by faculty, [AI] may improve accessibility' and another said AI could help with writing skills and 'level the playing field' for some students, including those for whom English is not their native tongue.'

Uncertainty of impact on educational equity

This section examines participants' four key takeaways uncertainty surrounding the impact of GenAI on educational equity, highlighting diverse perspectives, knowledge gaps, and concerns about the broader implications for higher education and society. The *context* of using AI in higher education and the concepts within a learning environment matter to how people responded to the impact on educational equity. The open-ended prompt did not clarify a positive or negative outcome, in which one faculty member said it was a 'vague question'. Data show that educational equity has various dimensions to it, and the 'definition of what 'education equity' means' with regard to this survey was not given. The impacts to educational equity 'will largely follow from how we as a university decide to implement and encourage/discourage its use'. Staff said it 'could help reduce gaps' and AI 'evens the playing field,' though what the person meant was unclear. Nonetheless, it was surprising how many responses indicated no knowledge of educational equity.

Uncertainty about how equity plays a role with AI echoed that of *context*, or the implications to society, higher education, teaching and learning, and systems. Nearly one-third of students were uncertain about the impact AI has on educational equity. People wrote 'no clue' and 'I don't know,' with several people mentioning it's too new of technology or that 'the jury is still out' to know the impact. Two staff members admitted to never thinking about the impact AI will have on educational equity and several others felt like they were 'not familiar with AI enough to determine its [sic] impact'. One student questioned the prompt and responded with 'it doesn't?' suggesting that AI might play a neutral or non-impacting role on educational equity.

Similarly, some written responses did not include equity and took a more general approach to AI, which we coded as N/A. For example, one faculty member said, 'creativity is now dead or at least moribund' and a staff member said, 'it can be used as a teaching tool for students'. Other comments suggest the institution needs to develop policies, that humans are still needed, and that AI can make education' become more interesting'. One staff member said they 'do not care about this

issue' and another 'I don't have an opinion on this'. Other responses were more directed to equality, than equity, such as GenAI 'evens the playing field' and it will 'provide more opportunity,' though it is unclear for whom the field is leveled, or opportunities given.

Several people mentioned their fears and concerns about GenAI technology as being larger than just education and that the implications will be felt throughout society. As 'it's not authentic or real at all' and:

> *Our sources get smaller and smaller as we use AI for anything. If AI use increases in the next 20 years, our 'updated' resources will come from a limited number of humans completing the research, or will research just stop and we reply on where we are now, and AI handles the updates? Concerned that research will be narrow and not creating more history. So, resources for faculty and students will be limited. [The Institution] needs to continue to put research, tracking, [and the] human aspect at the forefront.*

Not directed at educational equity, this staff person's response is calling out the potential for limiting resources by using AI. Similarly, one student questioned why they should pay for tuition if everything is already online, though how this impacts educational equity was left unstated.

Discussion and conclusion

Our results show that perceptions of equity related to GenAI in higher education are shaped by multiple factors of influence, extending beyond traditional concerns like biased content and physical access. While access remains an important issue, it is clear a more nuanced understanding of equity is needed—one that considers the broader context of skills, support systems, and institutional policies. Our findings highlight that equity concerns are not limited to who can afford or use GenAI tools but also include how these tools are applied and the role of literacy development in navigating them. Importantly, values and ethics are emerging as critical yet understudied dimensions of AI and equity. The diverse perspectives of students, faculty, and staff suggest that a singular focus on access and bias may obscure other essential factors that influence how GenAI tools shape educational equity.

While our student sample size was small (n=31) and not representative of the student population, the student responses were surprisingly focused on values, not access like the faculty and staff members. This finding was surprising, given the presumption of cheating frequency with GenAI tools (McMurtie, 2024; Wiley, 2024). This finding should not go unnoticed, as it appears integrity is important to them when using these tools; however, a discrepancy of what 'cheating' means may exist between students and faculty, adding to the confusion of what should be considered as academic dishonesty (Dahl & Waltzer, 2024). Eaton (2024) calls for moving past the traditional practices associated with academic integrity by decolonising what is considered acceptable academic output. By moving past these traditions, society will need to reevaluate what is 'ethical' and what is 'integrity' as our tools become more embedded in our future (Eaton, 2023). More research should be done with larger samples to see if and how 'values' remain a key aspect of students' GenAI use and students' definitions of cheating and academic dishonesty.

The dimension of access is crucial in revealing disparities in who can afford and effectively use GenAI tools. While faculty and staff tended to focus on physical access and financial barriers, Van Dijk (2017) critiques this view as narrow and argues the digital divide is often misunderstood as a simple gap between two groups. He explains that 'the metaphor suggests a simple distinction between two divided groups with a yawning gap between them,' which ignores the complexities and relative nature of digital inequalities (p. 1). Societal systems of oppression and privilege have been evident with digital access for decades as the importance of technology-literacy gained attention (Gorski, 2003). The early digital divide was rooted in physical access but neglected the

deeper, more nuanced issues related to skills, social support, and usage (Gorski, 2003; Van Dijk, 2017).

In our research, we see that this neglect persists in the rhetoric associated with AI equity. Van Dijk's notion of 'skills access' is particularly relevant here, referring to the 'digital and media literacy' skills necessary to use technology effectively (p. 7). Students who lack these skills are disadvantaged, even if they have physical access to the tools and will not resolve the divide (Gorski, 2003). A more multidimensional view of access, one that encompasses both the operational skills to command digital media and the substantive skills required to engage meaningfully with it, is needed. Additionally, Van Dijk highlights 'usage access,' where differences in how frequently and creatively individuals use technology further exacerbate inequities. Faculty who discourages students from exploring advanced uses of GenAI tools may inadvertently deepen socio-cultural inequities. Those with higher education tend to use digital tools for 'capital-enhancing goals' like career and academic growth (Van Dijk, 2017, p. 8), while others may only engage with simpler applications, such as entertainment (Bonfadelli, 2002; Cho et al., 2003; van Deursen & van Dijk, 2014; Van Dijk, 2017; Van Dyjk, 2005; Zillien & Hargittai, 2009).

As Van Dijk notes, barriers to access often go 'beyond' physical access to include 'computer anxiety' and 'technophobia,' which are particularly prevalent among underrepresented groups such as seniors, people with low education levels, and certain female populations (2017, p. 5). In our research, faculty, and staff express fears about biased GenAI content, but these concerns may overlook other critical barriers to equity, like the psychological and social obstacles that inhibit full engagement with these tools. The focus on content bias, though important, may obscure the more pressing need for skill-building and equitable support structures that empower all students to use and learn these tools meaningfully. When these skills go undeveloped, people are unable to gather and use information for informed decision-making, a form of information poverty (Marcella & Chowdhury, 2020), and an outcome of information inequity

In conclusion, while traditional concerns like physical access and biased content remain relevant, our findings reveal a deeper need to consider how factors such as skills, institutional support, and values influence equitable access and the use of AI tools. Equity is not merely about whether students can afford or physically access GenAI but also about the extent to which they possess the necessary skills, motivation, and opportunities to engage meaningfully with these technologies. To address the AI equity gap identified in this study, we propose that future research should prioritize the development of a multidimensional framework that integrates key equity dimensions such as access, skills, values, and institutional support. This framework should extend beyond traditional binary notions of access to encompass nuanced aspects like skills access, usage access, and literacy development as outlined by Van Dijk (2017). Additionally, it must consider the role of institutional policies in fostering an inclusive culture where diverse values and ethics are acknowledged and integrated into AI usage norms. Institutions should take a comprehensive approach that integrates AI literacy with an awareness of equity issues, fostering environments that actively support all members in overcoming barriers, whether they are technical, social, or psychological. A critical component of this framework would involve co-designing interventions with stakeholders—students, faculty, and staff—to ensure diverse perspectives shape strategies to bridge inequities. Such a framework could include mechanisms for fostering AI literacy through equitable support systems, embedding ethical AI usage in institutional curricula, and creating tools that mitigate psychological and social barriers like technophobia. By broadening the focus of equity discussions to include these dimensions, research would provide a robust foundation for developing policies and practices that promote meaningful engagement with GenAI across diverse educational contexts. Future research should expand on our findings, examining diverse institutional contexts and larger populations to deepen our understanding of how GenAI shapes educational equity and how we can address the potential gap.

About the authors

Sarah Zipf is a research project manager for Teaching and Learning with Technology at the Pennsylvania State University where she researches the implications and impact technology has on teaching and learning. Zipf's research interests include educational technology, distance and online education, educational equity, and outsourcing in higher education. She is an adjunct instructor for the Penn State World Campus. She can be contacted at stz2@psu.edu.

Chuhao Wu is a Ph.D. candidate in Informatics at the Pennsylvania State University. He holds a master's degree in Industrial Engineering from Purdue University. His research focuses on the human-centred design of educational technology and the application of artificial intelligence in higher education. He can be contacted at cjw6297@psu.edu.

Tiffany Petricini is an Associate Teaching Professor in the field of Communication Studies at Penn State Erie in Erie. She is co-chair of Penn State's Joint Standing Committee on Responsible and Effective Use of Artificial Intelligence in Higher Education, and co-chair of Penn State's Workload Policy Taskforce. She also leads the Penn State Artificial Intelligence Community of Practice and the Humanities Institute's Phenomenology Collaborative Colloquia. She also is the co-founder of the Educational AI Collaborative and is a Teaching with Technology Fellow for 2024-2026. Her publications have reflected interests in phenomenology, interpersonal communication, technology, philosophy, ethics, and media ecology. She also serves as the social media expert for NBC affiliate WFMJ 21 News. She can be contacted at tzr106@psu.edu.

Reference

Abrams, Z. (2024, April 1). Addressing equity and ethics in artificial intelligence. Monitor on Psychology. https://www.apa.org/monitor/2024/04/addressing-equity-ethics-artificial-intelligence

Amaral, A. (2022). Equity in Higher Education: Evidences, Policies and Practices. Setting the Scene. In O. Tavares, C. Sá, C. Sin, & A. Amaral (Eds.), Equity Policies in Global Higher Education (pp. 23–46). Springer International Publishing. https://doi.org/10.1007/978-3-030-69691-7_2

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. International Journal of Educational Technology in Higher Education, 21(1), 4. https://doi.org/10.1186/s41239-023-00436-z

Bonfadelli, H. (2002). The Internet and Knowledge Gaps: A Theoretical and Empirical Investigation. European Journal of Communication, 17(1), 65–84. https://doi.org/10.1177/0267323102017001607

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(May 2015), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brewer, J., Gerard, C., & Hakkinen, M. (2020, November 4). [EDUCAUSE Exchange]. https://er.educause.edu/podcasts/educause-exchange/the-impact-of-ai-on-accessibility

Cengage Group. (2023). 2023 Graduate Employability Report: AI Joins the Workforce (pp. 1–13). https://www.cengagegroup.com/news/press-releases/2023/cengage-group-employability-report/

Chang, K., Abdalla, Y. A., & Lasyoud, A. A. (2021). Artificial Intelligence in Personnel Management: Opportunities and Challenges to the Higher Education Sector (HES). In A. M. A. Musleh Al-Sartawi, A. Razzaque, & M. M. Kamal (Eds.), Artificial Intelligence Systems and the Internet of

Things in the Digital Era (Vol. 239, pp. 278–289). Springer International Publishing. https://doi.org/10.1007/978-3-030-77246-8_27

Cho, J., De Zuniga, H. G., Rojas, H., & Shah, D. V. (2003). Beyond access: The digital divide and Internet uses and gratifications. IT & Society, 1(4), 46–72.

D'Agostino, S. (2023, June 5). How AI Tools Both Help and Hinder Equity. Inside Higher Ed. https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2023/06/05/how-ai-tools-both-help-and-hinder-equity

Dankwa-Mullan, I. (2024). Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine. Preventing Chronic Disease, 21. https://doi.org/10.5888/pcd21.240245

Davidson, C. (2022). The new education: How to revolutionalize the university to prepare students for a world in flux (Updated version). Basic Books.

Gorski, P. C. (2009). Insisting on digital equity: Reframing the dominant discourse on multicultural education and technology. Urban Education, 44(3), 348–364. https://doi.org/10.1177/0042085908318712

Gorski, P. C. (2003). Privilege and repression in the digital era: Rethinking the sociopolitics of the digital divide. Race, Gender, & Class, 10(4), 145-176.

Grimes, S., & Feenberg, A. (2015). The SAGE Handbook of Digital Technology Research (S. Price, C. Jewitt, & B. Brown, Eds.). SAGE Publications Ltd. https://doi.org/10.4135/9781446282229

Heath, A. (2023, November 19). All the news from OpenAI's DevDay conference. The Verge. Alex

Humlum, A., & Vestergaard, E. (2024). The Adoption of ChatGPT. University of Chicago, Becker Friedman Institute for Economics Working Paper, 2024–50. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4807516

Jurado de Los Santos, P., Moreno-Guerrero, A.-J., Marín-Marín, J.-A., & Soler Costa, R. (2020). The term equity in education: A literature review with scientific mapping in web of science. International Journal of Environmental Research and Public Health, 17(10), 3526.

Kurban, C. F., & Şahin, M. (2024). Exploring ChatGPT's Role in Higher Education: A Literature Review. In The Impact of ChatGPT on Higher Education (pp. 41–73). Emerald Publishing Limited. https://doi.org/10.1108/978-1-83797-647-820241004

Levinson, M., Geron, T., & Brighouse, H. (2022). Conceptions of Educational Equity. AERA Open, 8, 233285842211213. https://doi.org/10.1177/23328584221121344

Lievrouw, L. A., & Farb, S. E. (2003a). Information and equity. Annual Review of Information Science, 37(1), Article 1.

Lievrouw, L. A., & Farb, S. E. (2003b). Information and equity. Annual Review of Information Science, 37(1), 499–540.

Luxton, D. D., & Watson, E. (2023). Psychological and Psychosocial Consequences of Super Disruptive A.I.: Public Health Implications and Recommendations. https://doi.org/10.25740/mg941vt9619

Lyerly, E. (2023). Utilizing ChatGPT to help students with disabilities. Disability Compliance for Higher Education, 28(9), Article 9. https://doi.org/10.1002/dhe.31479

Marcella, R., & Chowdhury, R. (2020) Eradicating information poverty: An agenda for research. *Journal of Librarianship and Information Sciences*, 52(2), 366-381. 10.1177/0961000618804589

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023). Artificial Intelligence Index Report 2023 (arXiv:2310.03715). arXiv. http://arxiv.org/abs/2310.03715

Petricini, T., Wu, C., & Zipf, S. T. (2023). Perceptions About Generative AI and ChatGPT Use by Faculty and College Students. EdArXiv. https://doi.org/10.35542/osf.io/jyma4

Porter, J. (2023, November 6). ChatGPT continues to be one of the fastest-growing services ever. The Verge. https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference

Quaye, S. J., Aho, R. E., Jacob, M. B., Domingue, A. D., Guido, F. M., Lange, A. C., Squire, D., & Stewart, D. L. (2018). A bold vision forward: A framework for the strategic imperative for racial justice and decolonization. ACPA—College Student Educators International.

Ratliff, K. A., & Smith, C. T. (2021). Lessons from two decades of Project Implicit. A Handbook of Research on Implicit Bias and Racism. APA Books. http://www.kateratliff.com/uploads/2/2/8/7/22879336/project_implicit_history.pdf

Rigano, C. (2018, October 8). Using Artificial Intelligence to Address Criminal Justice Needs. National Institute of Justice. https://nij.ojp.gov/topics/articles/using-artificial-intelligence-address-criminal-justice-needs

Sahota, N. (2024, January 15). AI For Equity: Bridging Racial Gaps With Artificial Intelligence. Forbes. https://www.forbes.com/sites/neilsahota/2024/01/15/ai-for-equity-bridging-racial-gaps-with-artificial-intelligence/

Schwarzman College of Computing. (2021, June 1). MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2021/potential-artificial-intelligence-bring-equity-health-care-0601

Sean Burns & Nicole Muscanell. (2024, April 15). EDUCAUSE QuickPoll Results: A Growing Need for Generative AI Strategy. EDUCAUSE Review. https://er.educause.edu/articles/2024/4/educause-quickpoll-results-a-growing-need-for-generative-ai-strategy

Suzman, M. (2023, May 21). AI Principles for Development to promote equity, Accelerate progress. Bill & Melinda Gates Foundation. https://www.gatesfoundation.org/ideas/articles/artificial-intelligence-ai-development-principles

U.S. Department of Commerce. (2023, October 31). How Artificial Intelligence Can Improve Diversity, Equity, Inclusion and Accessibility Efforts. U.S. Department of Commerce. https://www.commerce.gov/news/blog/2023/10/how-artificial-intelligence-can-improve-diversity-equity-inclusion-and

van Deursen, A. J., & van Dijk, J. A. (2014). The digital divide shifts to differences in usage. New Media & Society, 16(3), 507–526. https://doi.org/10.1177/1461444813487959

Van Dijk, J. (2017). Digital divide: Impact of access. The International Encyclopedia of Media Effects, 1, 1–11.

Van Dyjk, J. (2005). The deepening divide. Thousand Oaks. Sage.

Wanti, M., Wesselink, R., Biemans, H., & Brok, P. D. (2022). Determining factors of access and equity in higher education: A systematic review. Equity in Education & Society, 1(2), Article 2. https://doi.org/10.1177/27526461221092429

Wu, C., Zhang, H., & Carroll, J. M. (2024). AI Governance in Higher Education: Case Studies of Guidance at Big Ten Universities (arXiv:2409.02017). arXiv. http://arxiv.org/abs/2409.02017

Zillien, N., & Hargittai, E. (2009). Digital Distinction: Status - Specific Types of Internet Usage. Social Science Quarterly, 90(2), 274 – 291. https://doi.org/10.1111/j.1540-6237.2009.00617.x

Zipf, S., Petricini, T., & Wu, C. (2024). AI Monsters: An Application to Student and Faculty Knowledge and Perceptions of Generative AI. In The Role of Generative AI in the Communication Classroom (pp. 284–299). IGI Global. https://www.igi-global.com/chapter/ai-monsters/339072

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research – Vol. 45 No. iConf (2025)

Governing knowledge commons in information science

Ebubechukwu E. Uba and Madelyn Rose Sanfilippo

DOI: https://doi.org/10.47989/ir30iConf47281

Abstract

Introduction. The governing knowledge commons (GKC) framework supports analysis and research design regarding the co-production of information and communities, including associated governance mechanisms to structure engagement with the information or community produced in the interactions. However, knowledge commons are more prominent in information science (IS) as a concept than as robust theory to support analysis of commons, as through the GKC framework or other institutional approaches.

Analysis. This paper assesses the use of the GKC framework in IS research via structured and thematic analysis of the IS literature, collected systematically to include all published citations to GKC, the institutional analysis and development (IAD) framework, and commoning theory and references to knowledge commons.

Results. Results are organized to reflect the spectrum of engagement with these concepts and theories in the form of a typology. A guide is presented to apply the GKC framework for use in IS.

Conclusions. Robust theories support understanding and practice in IS research yet are often underutilized with respect to knowledge commons; future applications of the GKC framework will benefit research in knowledge management, information policy and ethics, and collaborative information and technology studies.

Introduction

Knowledge commons and institutional approaches to analysis of information policy and ethics, knowledge production, collaboration around information and technology topics are increasingly prevalent across interdisciplinary social science, including information sciences and informatics. Engagement with social theory in this process is scattered, reflecting a challenge for core information science (IS) objectives. IS focuses on data, information, knowledge, and wisdom throughout their lifecycles and as shaped by the people, contexts, and systems in which they are produced, stored, processed, shared, and used. IS aims to support both practice and deep understanding, often drawing on theoretical and methodological approaches from other disciplines. If engagement with theory regarding knowledge commons is superficial or sparse, this poses a challenge to understanding and practice. We aim to understand how knowledge commons theories and descriptive frameworks that support research design are used in IS and hope to offer guidance on best practices for this approach.

Analysis

A systematic review was conducted to understand: How is the concept of knowledge commons used in IS scholarship? What domains of IS engage with knowledge commons? When is the GKC framework or other robust institutional theory employed to support analysis of knowledge commons?

Data collection

Relevant literature was identified from both Web of Science (WoS) and Google Scholar. For WoS, a keyword search strategy identified relevant resources on knowledge commons (KC). This was further filtered by '*English language*' and filtered by research area '*library and information science*' to result in 25 articles. Considering only English-language publications, results are generalizable only to this realm of publication. More research articles were obtained via Google Scholar using the keyword '*knowledge commons*.' After reviewing the initial results, the terms '*information commons*' and '*learning commons*' were identified as related terms and they helped expand the search. Some of the articles were not related to the LIS research areas, the keywords '*knowledge commons in library and information science*' were used to refine the search results. Exclusion criteria were applied: articles that did not discuss knowledge commons or in relation to LIS were excluded. Keywords like '*information commons*' and '*learning commons*' were found to appropriate the concept of knowledge commons and were considered as distinct. This refined search strategy identified 20 additional research articles.

Content analysis

The full dataset of 45 relevant article were analysed via thematic, inductive coding of keywords and targeted, strategic qualitative analysis of uses of the concept knowledge commons, how citations to GKC approaches served the paper, and subdomains of IS represented in the paper. Coordination among the coders via discussion of coding for every reference ensured consistent application and analysis of themes (Richards & Hemphill, 2018). Iterative co-coding identified an array of uses of the concept of knowledge commons and levels of engagement with theory, resulting in classification of four distinct categories were identified: misuses of knowledge commons concepts; superficial engagement with knowledge commons concepts; analysis of knowledge commons without connection to existing theory; and meaningful theoretical analysis via knowledge commons theory.

Results

LIS subdomains relevant to knowledge commons

Multiple (library and information science) LIS subdomains engage with knowledge commons at some level, offering scope into where researchers might benefit from engagement with theory and clear conceptualization to ensure that scholars and practitioners understand their framing and its implications appropriately.

Library science and spaces as commons

The most common terms used as a substitute for knowledge commons in LIS were information commons and learning commons. Although, many of these references interpreted the terms as equivalents to knowledge commons (e.g., Lippincott & Skinner, 2022; Nwagwu & Matobako, 2022). This literature explores how libraries utilize information commons and learning commons as physical spaces, service models, learning and reading spaces, and spaces where rehabilitation sessions are conducted for students. It is important to acknowledge that '*information commons*' and '*learning commons*' have slightly different connotations within libraries and IS. Information commons often prioritize access to information technology and resources, while learning commons might focus more on collaborative learning spaces and pedagogical support.

Knowledge management and knowledge production

In knowledge commons discourse, the concepts of knowledge management and knowledge production were often explored. Knowledge commons was perceived as a means of producing and managing knowledge (e.g., Nishikawa, 2020). They can also serve as platforms for new knowledge creation through collaborative efforts and information exchange. This subdomain was referenced in studying the theories of knowledge management in one resource, and the need for enhance knowledge management to help mitigate the risk of information sharing.

Information policy and ethics

The policy and ethical considerations of knowledge commons is an important area of study. The resources surveyed in this literature raise issues of copyright and intellectual property, alternative rights regimes, and open access issues. Resources like those examined by Fisher and Fortmann (2010) were concerned with how data was being shared, owned, and governed within the knowledge commons, and, in this light, Peng (2013) discusses issues of open content licenses. They assume that since the knowledge commons is a shared resource and that there might be conflicts between what owns the contents and data shared. Knowledge commons theory provides a structural mechanism to disentangle these issues.

Collaborative information and technology studies

Studies of computer supported cooperative work and innovation studies address collaborative issues around information and technology. This subdomain raises questions related to open access. However, the knowledge resources examined herein are often less public or more microlevel knowledge commons as shared resources. These works emphasize the need for innovations to take place within knowledge commons. Concepts and terms like '*innovation-centric knowledge commons*', '*collaborative knowledge sharing*' emerge (Ramakrishnan et al., 2021). This subdomain focuses more on the technical aspects of collaboration and information technologies that might be used to support knowledge commons. The emphasis here is more on the social and organizational aspects of shared knowledge.

A Spectrum of engagement with knowledge commons

Across each of these domains, content analysis revealed a wide spectrum of engagement with knowledge commons as a concept and via social theory.

Misuses of knowledge commons

Various references demonstrate how knowledge commons, as shared information and resources that are collaboratively constructed, can be conceptually conflated with other concepts. These sources were so categorized often because they employed knowledge commons as synonymous with open resource, without social or collaborative sharing. For example, Nwagwu and Matobako (2022) apply the concept to describe a technology-equipped space in the library where people have access to digital resources and develop their digital literacy skills. This use of knowledge commons, rather than technology commons, is not unique to this study, but was prevalent in libraries and universities. While LIS as a field might recognize this as a distinct meaning of the concept, it is problematic when references are mistakenly cited with the opposite implication. Other resources were so classified because they treated knowledge commons as interchangeable with learning commons, with no clarity about how knowledge is contributed or accessed. For example, Raish and Fennewald (2016) interpreted knowledge commons as any learning space.

Superficial engagement with knowledge commons

Other resources did not engage with knowledge commons at an analytical level; these references considered knowledge commons without defining their meaning or analyzing the commons. Repositories as knowledge commons were studied (Nishikawa, 2020) to understand how research data were governed. Although repositories were identified as a knowledge common, there was no mention of a particular repository as a common. In parallel, Pang (2013) provides a brief understanding of knowledge commons, but analyzes primarily issues of copyright and licensing. Open licensing was suggested as an alternative, including via knowledge commons, but not explicated. Ferreira (2012) engages with knowledge commons at a surface level, by providing a good definition of knowledge commons, citing literature relative to how they function, and articulating key challenges, based on background literature. However, this article also did not analyze the commons within the data or consider deeper aspects like governance or power dynamics. Ferreira (2012) was an edge case for this category.

Analyzing knowledge commons without theoretical grounding

Some resources were categorized because they defined the concept of knowledge commons and could identify a knowledge commons but did not analyse it using institutional theory. Lippincott and Skinner (2022), for example, focused on analysis of the accountability in the Next Generation Library Publishing (NGLP), a knowledge-commons for scholarly publishing, using the FOREST framework for values-driven scholarly communication. The paper identifies the actors (scholars and scholarly communication stakeholders) and that governance principles that should be implemented but does not discuss them in detail. Kostakis and Papachristou (2014) discuss the potential of Commons-based Peer Production (CBPP) for 3D printing that enables local production in physical manufacturing. Though it does not engage substantially with theory, the authors argue that CBPP principles can enable collaborative design and empower local communities. Fisher and Fortmann (2010) employed a non-institutional theoretical framework to analyze FLUXNET, an example of a knowledge commons where scientific communities share data and control its governance. The paper, and its framework, do not detail the commons, but rather applies design principles from the common property theory (CPR) for successful commons management.

Theoretical analysis of knowledge commons

A subset of papers engaged with meaningful knowledge commons theory, drawing primarily on the institutional grammar (Crawford & Ostrom, 1995), the IAD framework, or the governing knowledge commons (GKC) framework. Notably, MacNaughton and Rao (2017) analyzed the Open and Collaborative Science in Development Network (OCSDNet) as a knowledge commons via the IAD framework, exploring new innovative mechanisms that can enhance collaborative disaster recovery planning, knowledge management, and learning in the Caribbean. These authors deeply explored action arenas, attributes of the common, and the governance mechanisms, grounded by

theory. Ramakrishnan, Shrestha, and Soar (2021) explore the connection between knowledge commons, innovation, and knowledge management (KM). They focus on proposing a new innovation-centric knowledge commons and analyze a conceptual model for Open Innovation Commons that utilizes KM principles.

Discussion

GKC (Frischmann et al., 2014; Sanfilippo et al., 2018) is at its core a descriptive framework to analyze information governance regarding contextual co-production of knowledge and community. GKC considers the social, economic, and political background in which information is produced and governed, as well as attributes of information and actors. The framework is informed by the longstanding political economic tradition of institutional analysis (Ostrom, 1990) and applies an institutional grammar (Crawford & Ostrom, 1995) to examine of strategies, norms, and rules that shape information throughout its lifecycle. Through a series of questions (Sanfilippo & Ocepek, 2024), it facilitates wholistic analysis of factors relevant to social knowledge production, sharing, use, and governance, as well as structured comparisons of individual supports and constraints within information governance (Figure 1).

Attributes	Aims	Conditions	STRATEGIES		
Attributes	Aims	Conditions	Modal Logic	NORMS	
Attributes	Aims	Conditions	Modal Logic	Consequences	RULES

Figure 1. An institutional grammar for knowledge governance

The GKC framework has structured diverse empirical research designs across many contexts, including health information (Strandburg et al., 2017), smart cities (Frischmann et al., 2023), and social media (e.g., Sanfilippo & Strandburg, 2021; Won & Klamer, 2021). The questions presented by Frischmann et al. (2014) can themselves define qualitative inquiry across contexts to collect or analyse data regarding data or information resources, social processes of knowledge production, and the efficacy or evolution of associated governance. Framework questions can be translated or applied to interview or survey questions (e.g., Sanfilippo & Strandburg, 2021) or adapted as a codebook for analysis of historical or governance documentation (e.g., Madison, 2021). Quantitative and mixed-method approaches are also supported by the institutional grammar and representative survey approaches (e.g., Shvartzshnaider et al., 2022). Further, given the structured nature of the framework, it can be meaningfully integrated with other social theory to explore specific types of information governance, such as privacy (Sanfilippo et al., 2020; Shvartzshnaider et al., 2022) and misinformation (Sanfilippo & Ocepek, 2024).

The GKC framework also supports rich conceptual analysis or results, by providing structure to compare case studies and understand the formation of new information norms or how actors produce information governance. As Figure 2 presents, the relational nature of categories of questions within the framework supports dynamic and longitudinal analysis of individual cases, as well as support comparisons across cases.

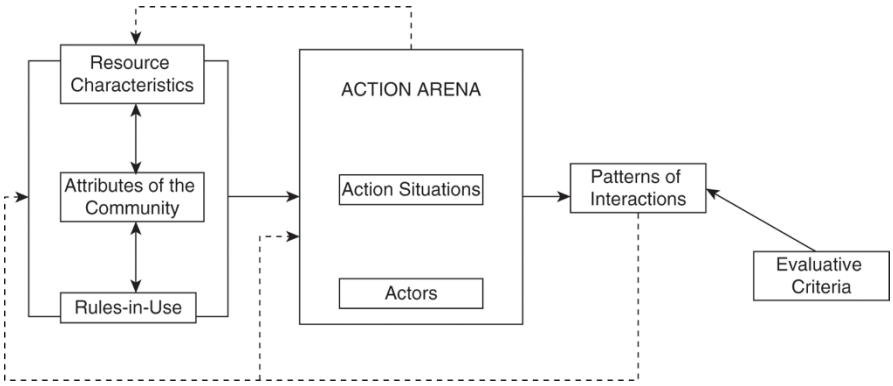


Figure 2. Processes within the GKC Framework

GKC approaches are pragmatic and iteratively developed over two decades of empirical and conceptual research to facilitate interdisciplinary research around social production, use, and sharing of information, core aims of wide swaths of IS research. For example, a current challenge to both KM and information policy subdomains of LIS is posed by generative AI and uncertainty people have regarding veracity and origin of information that influences decision-making, both among experts and everyday individuals. The GKC framework offers structure through which to explore how different communities might grapple with the issue in different ways, with some domains accepting AI-sourced content and others adopting constraints. Further, GKC studies of misinformation offer specific insights that are analogous, given the role of bots and deepfakes in various cases (Sanfilippo & Ocepek, 2024).

Conclusions

IS research benefits best support understanding and practice when supported by clear conceptualization and robust social theory. This conclusion briefly summarizes answers to our research questions. How is the concept of knowledge commons used in IS scholarship? Many subdomains of English-language IS, and LIS scholarship engage with knowledge commons concepts in order to explore shared knowledge resources, collaborative data and information governance and management, and open resources. What domains of IS engage with knowledge commons? Knowledge commons are explored relative to various domains, including LIS, information policy, KM, and collaboration. When is the GKC framework or other robust institutional theory employed to support analysis of knowledge commons? While many papers acknowledge with or engage rich knowledge commons theory, few fully apply them in their analysis. There is significantly more opportunity to apply the GKC theory and other institutional approaches. Future scholarship would benefit from deeper engagement with the GKC and the potential comparisons and practical policy and governance implications that can be translated through structured methods and analysis. Such an approach will better support refinement of knowledge commons in practice, as well as ensure consistent and clear use of the concept within the literature.

Acknowledgements

We appreciate the support of the school of information sciences at the University of Illinois at Urbana-Champaign and the NSF (Award # 2017495), as well as constructive feedback from reviewers.

About the authors

Madelyn Rose Sanfilippo is an Assistant Professor in the School of Information Sciences, University of Illinois, Urbana-Champaign USA. She received her Ph.D. from Indiana University, and her research focuses on governance in sociotechnical systems. She can be contacted at madelyns@illinois.edu

Ebubechukwu E. Uba is a PhD student in the School of Information Sciences, University of Illinois, Urbana-Champaign USA. Her research focuses on digital information access and literacy. She can be reached at euba2@illinois.edu

References

Crawford, S. E., & Ostrom, E. (1995). A grammar of institutions. American political science review, 89(3), 582-600.

Ferreira, S. (2012). Evolution and future of the knowledge commons: emerging opportunities and challenges for less developed societies. Knowledge Management for Development Journal, 8(2-3), 141-168. https://doi.org/10.1080/19474199.2012.717757

Fisher, J. B., & Fortmann, L. (2010). Governing the data commons: Policy, practice, and the advancement of science. Information & management, 47(4), 237-245.

Frischmann, B. M., Madison, M. J., & Sanfilippo, M. R. (Eds.). (2023). Governing smart cities as knowledge commons. Cambridge University Press.

Frischmann, B. M., Madison, M. J., & Strandburg, K. J. (Eds.). (2014). Governing knowledge commons. Oxford University Press.

Kostakis, V. & Papachristou, M. (2014). Commons-based peer production and digital fabrication: The case of a RepRap-based, Lego-built 3D printing-milling machine. Telematics and Informatics, 31(3), 434-443.

Lippincott, S., & Skinner, K. (2022). Accountability in the knowledge commons: lessons from Next Generation Library Publishing and the 2022 Values and Principles Summit. Insights, 35.

MacNaughton, M. & Rao, L. (2017). Governing Knowledge Commons: Applications of an Open Knowledge Broker in Caribbean Disaster Management. Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices, 245-256.

Nishikawa, K. (2020). How are research data governed at Japanese repositories? A knowledge commons perspective. ASLIB Journal of Information Management, 72(5); 837-852.

Nwagwu, W. E. & Matobako, M. M. (2022). Emerging knowledge commons in the public libraries in South Africa. Library Hi Tech.

Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge University Press.

Pang, N. (2013). Collaborative Production, the Knowledge Commons and the Application of Open Content Licenses. In Information Management Association (Ed.), Digital Rights Management: Concepts, Methodologies, Tools, and Applications (pp. 1162-1178). IGI Global.

Raish, V., & Fennewald, J. (2016). Embedded Managers in Informal Learning Spaces. Libraries and the Academy, 16(4), 793-815.

Ramakrishnan, M., Shrestha, A., & Soar, J. (2021). Innovation centric knowledge commons—A systematic literature review and conceptual model. Journal of Open Innovation: Technology, Market, and Complexity, 7(1), 35.

Richards, K. A. R., & Hemphill, M. A. (2018). A practical guide to collaborative qualitative data analysis. Journal of Teaching in Physical education, 37(2), 225-231.

Sanfilippo, M., Frischmann, B., & Standburg, K. (2018). Privacy as commons: Case evaluation through the governing knowledge commons framework. Journal of Information Policy, 8, 116-166.

Sanfilippo, M. R., & Ocepek, M. G. (eds.) (2024, in press). Governing Misinformation in Everyday Knowledge Commons. Cambridge University Press.

Sanfilippo, M. R., & Strandburg, K. J. (2021). Privacy governing knowledge in public Facebook groups for political activism. Information, Communication & Society, 24(7), 960-977.

Shvartzshnaider, Y., Sanfilippo, M. R., & Apthorpe, N. (2022). GKC-CI: A unifying framework for contextual norms and information governance. Journal of the Association for Information Science and Technology, 73(9), 1297-1313.

Strandburg, K. J., Frischmann, B. M., & Madison, M. J. (Eds.). (2017). Governing medical knowledge commons. Cambridge University Press.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Steering the AI world: an exploratory comparison of AI Acts in the EU and Canada

Ruiyi Zhu and Tien-I Tsai

DOI: https://doi.org/10.47989/ir30iConf47278

Abstract

Introduction. As artificial intelligence (AI) continues to grow rapidly, governments are implementing legislative frameworks to address its risks and opportunities. This paper provides a comparative analysis of AI Acts in the European Union (EU) and Canada, focusing on two legislative efforts: the EU artificial intelligence act (EU AI Act) and Canada's artificial intelligence and data act (AIDA).

Method. A summative approach of qualitative content analysis was used to examine the scope, risk classification, and regulatory strategies employed in the EU AI Act and Canada's AIDA. This study highlights similarities and differences in their approaches to managing AI's societal impacts.

Results. Both Acts provide positive directions and encourage responsible AI by addressing AI-related risks and opportunities. The analysis further explores the challenges, such as the definition of AI, enforcement mechanisms, and the inclusion of ethical considerations.

Conclusion. By drawing on these cases, the paper illustrates how regulatory steering can ensure responsible AI development and deployment in different geopolitical contexts. This paper offers insights into the evolving nature of AI governance and contributes to the broader discourse on balancing innovation with societal safeguards.

Introduction

Artificial intelligence (AI)-based technology has revolutionized industries and society, affecting the general public and the government by enabling advancements in various fields such as medical diagnosis, resume screening, and evaluating learning outcomes. Meanwhile, serving roles from filtrating existing data to generating new content, AI is also imposing a new challenge in the realm of information science. Due to its ability to efficiently collect, analyse, and visualise substantial personal information and use them to create new products through specialized algorithms, there is an ever-expanding demand for solid information management and protection to promote the ethical development of explainable AI, ensuring a higher level of data and privacy security. As a result, AI ethics, a field that '*studies the ethical principles, rules, guidelines, policies, and Acts that are related to AI*' (Siau & Wang, 2020, p. 74), has gained the attention of many scholars and stakeholders. In recent years, a number of released ethics guidelines '*comprise normative principles and recommendations aimed to harness the 'disruptive' potentials of new AI technologies*' (Hagendorff, 2020, p. 99), which led to extensive research on topics such as the perceived moral issues with AI, the features of an ethical AI, and to what extent the respective ethical values are implemented in the practice of application of AI systems.

However, some researchers also argue that

> *using ethics as a substitute for law risks its abuse and misuse. This significantly limits what ethics can achieve and is a great loss to the AI field and its impacts on individuals and society* (Rességuier & Rodrigues, 2020, p. 1).

suggesting the urgency of forming formal legislation in order to create unified and solid standards for AI use. As shown in prior research on policy, making comparisons is an essential tool for navigating the world, which should take into account the distinct development of different countries (Cagdas Artantas, 2023, p. 190). Therefore, studying AI-related laws of multiple jurisdictions is significant for providing us with a comprehensive overview of AI development within different social contexts. The research aims to answer the following questions: What are the general features of recently compiled AI Acts in the early-mid 2020s? Specifically, what are the countries' attitudes towards AI technology? What aspects of AI do the regulations focus on?

Method

The current study adopts a summative approach of qualitative content analysis, which involves a comparison of the original laws rather than using a formative approach to identify themes during the legislation process. The qualitative content analysis was conducted through the researchers' interpretation of the texts to identify the themes or patterns as suggested by Hsieh and Shannon (2005).

In order to set the criteria for AI regulation case selection, a wide variety of sources were consulted during mid-2024. Since AI is a rapidly evolving topic, the current study collected data after 2020. The primary sources, including both the original texts of the Acts and the companion document or Act summary available on the official website of the corresponding country, were found through Google Search by searching the title of each specific law. The secondary source and supplementary articles, including scholars' perceptions on the general concept of AI law and methodology of AI categorization, were found through keyword searches on several general databases such as the Web of Science and Scopus.

When identifying AI regulation cases for the current analysis, it can be noticed that most countries express concerns regarding AI by addressing it in regulations from specific aspects, such as digital platform transparency, environmental impact, risk management, and financial instruments and exchange (Anand, 2024; Kamiya & Keate, 2024). As of mid-2024, many countries are still at the conceiving step of creating their legislation, while some are relatively conservative and are waiting to draw from others' experiences. For example, the US currently has compiled only a *Blueprint for an AI Bill of Rights*. Although it has state-level legislation, its feasibility is limited as compliance

issues can arise from overlapping or conflicting regulations (Plotinsky & Cinelli, 2024). Nations including the UK, Mexico, and Chile have introduced *artificial intelligence (Regulation) bills*, which are still awaiting approval from the legislatures and executives before coming into effect as law. These regulations are not included in this research because such documents are still subjected to further modifications, and thus, they can neither stand for a country's final decision on whether or not to create the federal AI legislation nor provide a reliable date of entry into force. Similarly, due to modern debate revolving around the definition and classification of AI, Japan, like the US and the UK, also chose to '*respect companies' voluntary governance and provide non-binding guidelines to support it, while imposing transparency obligations on some large digital platforms*' (Habuka, 2023). Additionally, China has published an *interim regulation on the management of generative artificial intelligence services* that entered into force in August 2023. Nevertheless, the nature of an interim regulation that takes effect immediately may not be the same as an act. The paper starts the discussion by examining AI Acts that have followed the regular development procedure; thus, only AI Acts were included as cases of investigation.

According to the above navigation and considerations, the selection scope of the current study includes AI Acts published in the 2020s and the scale of the Acts at the federal level, focusing only on federal-level legislation that has already been published or is scheduled to enter into force by the end of 2025, of which original texts can be accessed on the official government websites, which means they have a certain level of completeness and have been reviewed by the executive for at least multiple times. The cases selected for the current study are the EU artificial intelligence Act from the European Union and the artificial intelligence and data Act (AIDA) from Canada.

Based on Habuka's (2023) classification of AI regulations, two general approaches towards AI regulations include: (1) a '*sector-specific and soft-law-based*' approach that applies to countries like the US, the UK, and Japan, and (2) a '*holistic and hard-law-based*' approach that has been taken by countries and political unions like EU and Canada. The selected cases in this study fall into the latter category, characterized by their capability of applying to all AI applications across all sectors and applications, as well as having regulative authority that lies with the central government (Hilliard, 2022).

Findings and discussion

Overview of AI Acts

Table 1 provides a summary of the basic features of the AI Acts, which are intended to be used to compare general information, such as the types of AI being addressed, the objectives of the laws, the affected parties, and the implementation methods. The data are summarized from the government websites and texts of the original Acts, thus standing only for how the laws are expected to function in ideal situations. Since there are limited records or studies on the laws' effectiveness and defects in practice, their real-life performances are not considered here.

	European Union	Canada
Name of Legislation	EU Artificial Intelligence Act	The Artificial Intelligence and Data Act (AIDA)
Governmental Agencies Responsible for Implementing the Acts	• The AI Office will be established, sitting within the Commission.	• Minister of Innovation, Science, and Industry • AI and Data Commissioner
# of Clauses/Chapters	13 Chapters	41 Clauses
Types of AI's /being Addressed	High-Risk AI Systems General Purpose AI (GPAI)	High-Impact AI Systems
Major Objectives	• Address risks to health, safety, and fundamental rights. • Protects democracy, rule of law, and the environment. • Create a balanced regulatory environment that fosters innovation.	• Reassure Canadians that the government has a thoughtful plan to manage the emerging technology and maintain trust in a growing area of the economy. • Reassure actors in the AI ecosystem in Canada that the aim of the Act is not to entrap good faith actors or to chill innovation, but to regulate the most powerful uses of this technology that pose the risk of harm.
Parties Responsible for Complying the Acts	• Providers of high-risk AI systems • Users (deployers) of high-risk AI systems have some obligations, though less than providers (developers). • Providers of GPAI (General Purpose AI) models	Businesses who: • Design or develop a high-impact AI system • Make a high-impact AI system available for use • Manage the operation of an AI system
Examples of Measures being Implemented	• Establish a risk management system. • Conduct data governance, ensuring that training, validation, and testing datasets are relevant and representative. • Provide instructions for use to downstream deployers. • Design their high-risk AI system to allow deployers to implement human oversight.	• Identify and address risks with regard to harm and bias. • Document appropriate use and limitations. • Adjust the measures as needed. • Ensure ongoing monitoring of the system.
Penalties	• Warnings and Non-Monetary Measures • Monetary Measures	• Administrative monetary penalties (AMPs) • Regulatory Offences

	European Union	Canada
Name of Legislation	EU Artificial Intelligence Act	The Artificial Intelligence and Data Act (AIDA)
Governmental Agencies Responsible for Implementing the Acts	• The AI Office will be established, sitting within the Commission.	• Minister of Innovation, Science, and Industry • AI and Data Commissioner
# of Clauses/Chapters	13 Chapters	41 Clauses
Types of AI's /being Addressed	High-Risk AI Systems General Purpose AI (GPAI)	High-Impact AI Systems
		• True Criminal Offences
Date of Entry into Force	• August 1st, 2024. • After entry into force, the AI Act will apply by the following deadlines: • Six months for prohibited AI systems. • 12 months for GPAI. • 24 months for high-risk AI systems under Annex III. • 36 months for high-risk AI systems under Annex I.	• No sooner than 2025.

Table 1. An overview of AI Acts in EU and Canada (Sources: EU (2024) and AIDA (2003))

The latest versions of both AI Acts were published between 2023 and mid-2024, which marks the great importance attached to this modern technology. Upon comparing the general features of both Acts, a high acceptability of AI development with a positive attitude toward AI can be identified. To be geared to increasing international competitiveness and productivity, governments are guiding AI innovation in a positive direction and encouraging the responsible adoption of AI technologies. Under such context, the created Acts share some similar features which are introduced and discussed in the following sections.

Holding a positive attitude toward AI

The AI Acts generally show a positive attitude toward AI's potential to facilitate the evolution of new-age technology. Recognizing the unique nature of the AI ecosystem, the laws indicate the motivation of adapting to the shifting landscape by encouraging exploring the optimization of usage scenarios and implying the tendency to cooperate with different parties—like industry associations, enterprises, education, and research institutions, public cultural bodies, and relevant professional bodies— in areas such as the establishment of data resources, applications, and risk prevention. On the other hand, the regulations are relatively conservative about enforcing punishment. For example, when implementing monetary penalties, many clauses use general phrases such as '*not more than...*' (AIDA, 2023) or '*up to...*' (EU AI Act, 2024) a certain amount rather than specifying the punishment standards for different situations, while others need to refer to

existing laws and administrative regulations for detailed solutions (Management of Generative AI, 2023).

Focusing on AI developers rather than end-users

Upon considering the influence levels different parties have when interacting with AI, the regulations display another noticeable trend—both Acts focus on the behavioral accountability of AI developers and deployers rather than end-users. This decision may need to work together with the idea of AI ethics, as without relevant rules, users are expected to self-consciously use AI in proper contexts by following the moral principles and guidelines after the AI developers state any restrictions on how the systems are meant to be used and their limitations. While it is reasonable to focus on AI developers at this stage, the rights and responsibilities of end-users may also be important issues to be considered in the future.

Ensuring fairness and inclusivity

A common concern expressed by AI Acts is the prejudice related to information screening as they all include clauses that touch on a similar matter—during processes such as algorithm design, the selection of training data, model generation, and optimization, and the provision of services, effective measures should be employed to prevent the creation of discrimination such as by race, ethnicity, faith, nationality, region, gender, age, profession, or health. Potential biased output occurs when there is an unjustified and adverse differential impact based on unfair data selection, which is also considered one of the main sources of physical harm, psychological harm, damage to property, and economic loss to an individual or broadly across groups. For example, the past analysis of a well-known facial recognition system showed evidence of bias against women and people of color (Buolamwini et al., 2018), which reveals how the protection of human rights becomes an urgent issue as technology may learn from the mistakes in past data and repeat it unconsciously.

Classifying AI Systems based on its risks

Considering the lack of a universal classification of AI impact and risk, the EU and Canada propose their own assessment standards in the regulations. When determining which AI are high-impact, AIDA considers particularly systems that affect access to employment, influence human behavior at scale, are critical to health and safety, and are used for identification and inference (AIDA Companion document, 2023). Meanwhile, the EU AI Act defines high-risk AI systems as those used as a safety component or a product covered by specific EU laws and those applied to fields including non-banned biometrics, critical infrastructure, education and vocational training, employment, and workers management, etc. (Summary of the EU AI Act, 2024). It can be seen how the addressed AI technologies reach far beyond narrow AI systems, which rely on specific algorithms to perform simple tasks within a limited domain like recognition or translation. High-impact and high-risk AI systems thereby possess the capability to affect further human decisions and administration processes in different fields. Overall, generative AI shows similar characteristics to the high-impact and high-risk AI mentioned in the Canadian and EU AI Acts by having significant influences on public information and privacy security.

Unclear and unaddressed elements

Despite the regulations showcasing the effort of limiting harmful behaviors and promoting systematical AI development, there are several elements that have been neglected and require further clarification.

The scopes of the AI Acts

While the laws clearly apply to AI developers and servers in relevant industries and fields, it is ambiguous to what extent they can restrict AI use in government institutions, which problem contradicts the objective of the regulations. Stating the specific type of systems that will be used by the federal department can help improve the transparency of the legal code implementation

process, which is crucial for showing the public a trustworthy regulatory environment and avoiding variation in AI act standards or deviation from international accountability measures.

The definition of AI

When analyzing Japan's AI regulation, Kamiya and Keate (2024) point out that there is no legally recognized definition of 'AI' in Japan, and it is difficult to define its scope strictly *because 'AI is an abstract concept that includes AI systems themselves as well as machine-learning software and programs.*' The same predicament applies to other nations as well since the definition of this terminology is inconsistent across fields. This issue may pose challenges by leading to situations such as when '*the same AI systems may fall under one law but not the other, potentially limiting individuals' recourse for rights violations*' (Muhammad & Yow, 2023, p. 512). As a result, establishing a universal definition that addresses fundamental concerns and specific techniques associated with AI is essential to ensure coherence and hold relevant organizations accountable.

Conclusion

While the development path of AI is still unpredictable, the emergence of federal-level regulations in recent years reveals some methods different nations take towards steering innovative yet unstable technology by ensuring human rights are thoroughly considered and protected at each stage of the lifecycle of an AI system. The '*holistic and hard-law-based*' approach taken by the EU and Canada through '*[setting] forth obligations—such as for governance, transparency, and security for at least high-risk AI—with significant sanctions in case of violation*' (Habuka, 2023), is able to manage all types of business simultaneously with the same set of rules. Meanwhile, when keep improving the soundness of the law, legislators should also establish measures to prepare societies for the upcoming challenges in the near future, by when '*issues such as whether algorithms could be protected by free speech rights and artifacts produced by AI that could be protected by copyright laws*' (Jia & Zhang, 2022, p. 60) may become typical examples. The close coordination between AI legislation and AI ethics theory is indispensable for preparing the public for the dynamic change during the age of technology evolution.

Overall, the current study aims to initiate the conversation based on a few selective cases and provide a preliminary framework to explore and examine AI Acts. Given the rapid advancement of AI technology and the growing recognition of the need for effective AI regulations, it is anticipated that more AI Acts will soon be developed and enacted. For instance, after we finished the current data analysis, Brazil officially approved its bill in December 2024, establishing a defined legal framework for the nation's AI use. Future research should expand the scope to encompass a wider range of AI regulations for more comprehensive analysis. While AI Acts may further develop to clarify the scope and the definition as well as address more on the rights and responsibilities of different stakeholders (including end-users), AI Acts may need to retain flexibility so it can adapt to technological advancement. This adaptability may help foster an evolving AI governance framework that ensures the responsible and sustainable development of AI.

Acknowledgements

The authors are thankful for the support from the National Science and Technology Council, Taiwan (NSTC 112-2410-H-002-015-SS2).

About the authors

Ruiyi Zhu is an undergraduate student at the University of British Columbia, BC, Canada. She can be contacted at zry0826@student.ubc.ca

Tien-I Tsai is an Associate Professor in the Department of Library and Information Science, National Taiwan University, Taipei, Taiwan. She can be contacted at titsai@ntu.edu.tw

References

Anand, N. (2024, March 8). US Federal AI Legislation in 2024: The Current Landscape. Holistic AI. https://www.holisticai.com/blog/us-federal-ai-legislations

Banafa, A. (2024). Generative AI and other types of AI. Transformative AI (1st ed.), pp. 33-39. River Publishers. https://doi.org/10.1201/9781032669182-7

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1–15. Allen, Kate.

Cagdas Artantas, O. (2023). Promotion of green electricity in Germany and Turkey: A comparison with reference to the WTO and EU law (1st 2023.;1;1st 2023; ed.). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44760-0

Government of Canada. The Artificial Intelligence and Data Act (AIDA) – Companion document. (2023, March 13). Retrieved from https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

Habuka, H. (2023, February 14). Japan's Approach to AI Regulation and Its Impact on the 2023 G7 Presidency. Center for strategic & international studies. https://www.csis.org/analysis/japans-approach-ai-regulation-and-its-impact-2023-g7-presidency

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines (Dordrecht), 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

High-level summary of the AI Act. (2024, February 27). EU Artificial Intelligence Act. Retrieved from https://artificialintelligenceact.eu/high-level-summary/#weglot_switcher

Hilliard, A. (2022, August 6). Regulating AI: The Horizontal vs Vertical Approach. Holistic AI. https://www.holisticai.com/blog/regulating-ai-the-horizontal-vs-vertical-approach

Hsieh, H.-F., & Shannon, S.E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277-1288.

Interim Measures for the Management of Generative Artificial Intelligence Services. (2023, July 10). The State Council the People's Republic of China. Retrieved from https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm

Jia, K., & Zhang, N. (2022). Categorization and eccentricity of AI risks: A comparative study of the global AI guidelines. Electronic Markets, 32(1), 59-71. https://doi.org/10.1007/s12525-021-00480-5

Kamiya, M., & Keate, J. (2024, July 1). AI Watch: Global regulatory tracker - Japan. WHITE & CASE. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-japan#:~:text=As%20noted%20above,%20at%20this,and%20principles%20announced%20in%20Japan.

Muhammad, A. E., & Yow, K. (2023). Demystifying Canada's Artificial Intelligence and Data Act (AIDA): The good, the bad and the unclear elements. Paper presented at the 510-515. https://doi.org/10.1109/CCECE58730.2023.10288878

Plotinsky, D., & Cinelli, G. M. (2024, April 9). Existing and proposed federal AI regulation in the United States. Morgan Lewis. https://www.morganlewis.com/pubs/2024/04/existing-and-proposed-federal-ai-regulation-in-the-united-states

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, 7(2), 205395172094254. https://doi.org/10.1177/2053951720942541

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management*, *31*(2), 74-87. https://doi.org/10.4018/JDM.2020040105

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Open access: Māori perspectives on publishing

Spencer Lilley

DOI: https://doi.org/10.47989/ir30iConf47275

Abstract

Introduction. This short paper presents the results of a project that looked at Māori researchers' attitudes to open access publishing. It presents an overview of their engagement with green and gold publishing options, including barriers and challenges they encountered when considering open access publishing and the role that Māori liaison librarians have in encouraging Māori academics to utilize the range of open access resource opportunities available.

Method. In this study, six Māori researchers and four Māori liaison librarians were interviewed using a kaupapa Māori methodological approach.

Results. The results were analysed using a tikanga Māori focused lens

Conclusions. Although the Māori researchers in this study are aware of open access publishing, they were unsure of how to distinguish between the different models on offer and would benefit from regular engagement and advice from Māori liaison librarians.

Introduction

Scholarly publishing traditions have changed quite dramatically over the last 30 years, with the dependence on print-based journals gradually shifting to hybrid publication (Print and Online), and online only access. Pinfield et al. (2020) identify the beginning of the 21st century as the turning point for open-access publishing. Advocacy for open-access publishing has also resulted in publishers changing their business models, including the introduction of Article Processing Charges (APC), which shift the cost of publishing to academic authors. The costs associated with these can extend into five figures if an author wishes to make their article available as gold open access in a highly ranked journal, for example, gold open access for *Nature* (2024) is £8890/ $12290/ €10290. However, other journals provide different choices for their authors, including freely open access or an option to load a pre-print copy of manuscripts to institutional repositories. In response to changing market conditions, major publishers are entering into read and publish agreements. New Zealand universities have access to these as consortium members of the Council of Australian University Librarians (CAUL). These agreements enable university researchers the opportunity to publish their research outputs in open-access journals without having to pay an APC. There are also growing expectations that scholarly outputs from publicly funded research in New Zealand will be made openly available as a way of ensuring equity of access to these materials. While in addition to open access publishing, there is increased advocacy for researchers to deposit their data in open data repositories to make it discoverable and useable, this is currently not a requirement for publicly funded projects.

Based on empirical research, this paper focuses on the attitudes of researchers and librarians identifying as Māori (Indigenous peoples of New Zealand) to open-access publishing. The paper also considers the ethical issues associated with making Indigenous knowledge freely available via open access and open data platforms.

Background

Māori academics employed in New Zealand universities are expected to produce high-quality research outputs in nationally and internationally ranked journals. Although currently suspended, university researchers over the last two decades have been required to participate in the performance-based research fund (PBRF) evaluation exercise. This exercise awarded a quality score for each researcher which was measured against criteria to determine the contribution their research was making to creating leading-edge knowledge. While the research sector waits for the PBRF to be resumed or for the introduction of a new form of research assessment, researchers continue to have pressure placed on them by their institutions to publish in highly ranked publications. This pressure seriously impacts where researchers decide to publish or make their research available. From the point of view of their institution, a researcher should be aiming for publishers and journal titles which are considered to be highly ranked and prestigious, as this will greatly increase the reputation of the university and the individual. However, for a Māori researcher who has a strong commitment to kaupapa Māori (Māori methodological and theoretical approach) research principles, this is at odds with their sense of responsibility to the communities they are working with, who should be the primary beneficiaries of the research they have participated in and contributed to. Open access publishing provides researchers with the opportunity to make their research freely available, while also complying with the wishes of their institution. However, a recent report from the Council of New Zealand University Librarians (CONZUL, 2022), revealed that a significant majority of New Zealand research remains behind publisher paywalls. Country rankings from Scimago show that in 2022 their assessment of documents from New Zealand showed that only 52.18% of these were available as open access, although, as shown in Table 1, this is a vast improvement from the results in 2017. The CONZUL (2022) report stated that if all eligible manuscripts produced by New Zealand researchers in 2019 were deposited in repositories, the open access rate for that year would have increased to 70%.

Year	**Number of documents: Open Access**	**Percentage of annual outputs**	**Number of documents: Not Open Access**	**Percentage of annual outputs**
2017	7073	41%	10177	59%
2018	7719	43.25%	10127	56.75%
2019	8332	45.29%	10065	54.71%
2020	8959	48.81%	9396	51.19%
2021	9768	50.88%	9429	49.12%
2022	10117	52.18%	9272	47.82%

Table 1. Number and percentage of open access and not open access New Zealand outputs

Literature review

There is a paucity of literature related to how Māori researchers engage with the open-access publishing environment, with these issues only being covered in reports in grey literature published by CONZUL (2021, 2022), Saunders (2022), Henk (2019), and White and Remy (2017). These reports reveal low levels of understanding and uptake of open access publishing opportunities, with a reluctance to use gold open access routes and limited understanding of green open access options. The reports demonstrate that these practices are not only restricted to Māori researchers. The literature on Māori experiences of academic publishing reveals the impact New Zealand's PBRF research assessment process has had on Māori research. Roa et al. (2009) identified fourteen problems stemming from the PBRF process that had impacted on Māori research, with seven of these relating to the publication, dissemination and peer recognition aspects of the evaluation exercise and the academic or research behaviour expected of Māori researchers to obtain higher quality scores. The authors' concerns centred on the pressure that Māori researchers were under to produce scholarly outputs, which weren't always going to be of benefit to those who were subjects of the research but were within the '*publish or perish*' mentality existing in international university systems. This is a similar concern to that expressed by Tawhai, Pihera, and Bruce-Ferguson (2004), who were concerned that the PBRF assessment gave too much weight to the degree of international recognition kaupapa Māori research received, rather than the value it had to local iwi (tribal community) who commissioned or participated in the research project.

The wider international literature on Indigenous researchers' experiences of open access publishing demonstrates that the pressure to publish is also impacting Indigenous scholars. Bryant and Thomas (2024) reveal in their study that black, indigenous and people of colour (BIPOC) faculty have become more aware of the open access opportunities available to them but are still facing challenges due to exclusionary practices at their institutions. Berger (2021) and Masaka (2018), both point towards the use of open access as another way Western research dominates and obscures the impact of Indigenous research in Africa and other parts of the global south. Smith et al. (2021) confirm the dominance of open access publishing by authors from high-income countries, with APC costs being a barrier for researchers from the global south.

Methodology

A kaupapa Māori approach was utilized to understand how Māori researchers and Māori subject librarians engage with the scholarly publication process, identify how and where researchers decide to publish, and consider the role that Māori librarians have in advising them about the options available, including in the open publishing space. Kaupapa Māori methods are emancipatory, having been derived from critical theory, and created to provide Māori with the ability to assert their self-determination and demonstrate the validity and legitimacy of Māori traditional knowledge and customs (Smith, 2021). These methods enable participants and researchers to form a strong and trusting relationship based on tikanga Māori (Māori values and

beliefs), mātauranga Māori (Māori cultural knowledge) and a commitment to analysing the resulting data through a lens that is Māori focused. Other critical aspects of kaupapa Māori research include the development of mutual trust, respect, and reciprocity by ensuring that the community is the primary beneficiary from the outcomes. These requirements can at times restrict how and where the researcher is able to share their scholarly outputs from the research, as the community might be reluctant to have the knowledge they have shared disseminated into the public domain.

Participants

Six researchers and four librarians who identify as Māori were recruited for this project, which is part of a larger project on academic publishing and peer review experiences of Māori researchers. All ten participants are employed in New Zealand universities. The researchers included representatives from early (up to 7 years post PhD), mid (from 8 to 15) and late (15 years and over) career positions. All six were researchers, utilising kaupapa Māori research methodologies, actively publishing in New Zealand and international publications. The researchers came from a mix of disciplines including management, education, history, Māori studies, psychology, and science. The librarians were all highly experienced professionals, responsible for delivering services and resources to university staff and students, including advising researchers about where and how to publish their scholarly outputs.

Following recruitment, each participant was interviewed for up to one hour. Due to restrictions resulting from the COVID-19 pandemic, only five of the interviews were conducted in person, with Zoom being used for the other five. The early part of the interview was spent on whanaungatanga (relationship building), which entails the interviewer and participants establishing connections and understandings based on cultural and professional experiences. A core part of this process is to identify how the interview is going to be conducted, what use (if any) of Māori protocol will be followed and reiterate the participant's rights in keeping with the university's ethical code. As the interviews involved academics and librarians, a different range of questions was applied to each group. The interviews with academics involved asking participants six open-ended questions aimed at gaining a detailed understanding of their experiences of publishing their research in open and non-open contexts, including their perspective on the advantages and disadvantages of open access journals, and their perceived barriers to publishing in both contexts. The questions for librarians focused on the interactions they have with Māori researchers, including the advice they provide to researchers about publishing, and their perceptions of the awareness Māori researchers have about open access options available to them and different categories that open access consists of.

Results

The interviews with Māori researchers demonstrated that they all had an awareness of the push to open-access publishing and the contribution it is making in their discipline. However, the results revealed some confusion about the options that open-access publishing provided. There was also a perception that some fully open-access journals were of an inferior quality to those published in print or using a hybrid method. This was voiced as a view that the peer review processes were sometimes not as robust in new open-access journals as those in journals with a longer publishing history. Two of the participants mentioned that they were receiving multiple invitations every week from journals seeking articles or offering them the opportunity to be a guest editor of a special issue. Most of the invitations were from journals that required relatively high APCs. One of these researchers indicated that they checked regularly with library staff to ascertain the validity of their concerns about these journals. The more experienced researchers talked about how the academic publishing landscape had changed over the past decade, from being predominantly print based where there was a substantial delay between items being accepted for publication and their appearance in a print journal, to a situation where their articles were now made available 'online first'; they were mainly happy with this. Questioned further about their understanding of the

differences between diamond, gold, and green options of publishing open access, only two had heard about diamond, while all six could distinguish between gold and green publishing. However, the cost of publishing via gold was seen as a significant barrier, and there was a reluctance to pay an APC unless they had access to funds from their institution or an external funder to cover these costs.

Researchers were all very conscious of the need to ensure that their community participants benefit from their participation in the research, as this is one of the key components of doing research with Māori, and a failure to do this would hinder future research relationships with their communities and a loss of mana (status). However, all were equally aware of the need to be published in highly ranked journals so they could meet the quality expectations of their institutions. This need was also linked to improving their academic promotion prospects, which meant they would target a journal with a high ranking and worry about issues like providing open access afterwards. Consistently publishing in high-ranking publications was seen as a personal achievement and this would see their mana within academia grow immensely, but it was felt that this came after ensuring they had contributed back to the communities involved in their research.

As journal articles are written for an academic audience, there was also a belief they were not always the best way of communicating with their participants, and researchers would find alternative means of reporting back to these communities. One indicated that whenever possible they would organise hui (meetings) with her communities, so they could provide a summary of the results of the research and work with those present to identify how they could use the results for future advancement.

All six researchers were aware of the option of loading items into institutional repositories. However, it was clear that not all the researchers understood the rules or requirements of publishers related to Green open access and depositing pre-print (before peer review) or post-print (after peer review but before journal formatting occurs). There was also low awareness of the read and publish arrangements that have been negotiated with publishers, with only two of the researchers identifying this as an option they had taken advantage of.

All participants indicated that they liked to publish in New Zealand focused journals when appropriate, but because not all of these were open access, available online or indexed by citation databases like Scopus or Web of Science (affecting ranking), these would not always be their first choice.

Although not a specific focus of this project, the topic of open data was brought up by two of the participants. They mentioned there was increasing pressure for them to make the data they collected available for other researchers to access. They thought this trend was contrary to the principles of accountability they had to their communities, and raised broader issues associated with Māori data sovereignty and protecting data from re-use or misappropriation without the source community being consulted. They also thought it would be difficult to get participants involved in research if this requirement was enforced by journals or scholarly organisations.

Data from the interviews with Māori librarians related to open access issues revealed their willingness to work collaboratively with researchers wanting to identify where and how to publish. They perceived that although their skills and advice were called on by some researchers, many others would benefit from seeking their assistance. It was noted that as Māori liaison librarians they provided services to Māori researchers across many disciplines, so they needed to be highly aware of the complexities of the publishing process across subject areas. Where necessary they would call on the knowledge of other liaison librarians to assist Māori researchers needing specific advice related to their area of expertise. This collaboration often extended to the provision of special information literacy sessions focused on identifying potential publishing options. However, the uptake of these sessions by Māori researchers was often quite low, with an added observation

that advice was normally sought by individual researchers as and when required and not in advance. Although they also perceived that there was an increasing interest in open access options from Māori researchers, uptake was slow but steady. They believed there was much more that could be deposited in the online repositories than was currently the case. The librarians believed this was partly due to misunderstandings about publisher-imposed restrictions and what is permitted, and although they communicated with their academic communities regularly and information is available on their library's website, they were unsure of their impact on increasing use of the library repositories. In all cases, the librarians were also responsible for providing advice about the research process to students doing master's and doctoral research. This advice extended to publishing opportunities, and this was of particular interest to doctoral students wishing to publish articles as the basis for their dissertations. These students were predominantly concerned about the peer review processes and quality rankings of journals, preferring publications with strong reputations and not always understanding there were open-access options available to them that would enable them to make their work more accessible, while not compromising their wishes for robust peer review processes.

Discussion

Although acknowledging its importance to scholarly communication, it is clear from the interviews that open-access publishing by Māori researchers is still in the development phase. There is an obvious need for more education about the range of open access opportunities available, including how to make the best use of green access options. The cost of gold access is still regarded as a barrier due to low levels of funding being available for this purpose. However, the read and publish agreements that have been negotiated with publishers is an underutilised option taken up by Māori researchers, and librarians need to persist in educating researchers about its benefits. Librarians can also provide additional guidance about the quality of open access-only journals, including the robustness of the peer review processes. It is clear that Māori researchers are conscious of the need for their research to be of primary benefit to their participants, and although open access provides these individuals with free access to articles, there is a recognition that there might be more appropriate methods to communicate research outcomes with these communities. Although there was little focus on issues related to open data, it is an aspect of the research landscape that is becoming more important. Māori researchers in disciplines that expect research data to be openly available will need to consider how these expectations relate to concerns around Māori data sovereignty and the reuse of data without the source community's informed consent.

Conclusion

Although it is a small study and its results are not generalisable, this paper has demonstrated that there is an increasing focus on open-access publishing within New Zealand's academic institutions. The evidence from this study and the grey literature shows that the uptake of these opportunities has slowly grown over the last seven years. However, it has also shown that although there is interest in making their research more freely accessible, the Māori researchers in this study are not fully aware of how they can take advantage of open access opportunities available to them. This situation could be improved through greater engagement with the services and advice offered by Māori liaison librarians, providing Māori researchers with increased opportunities to fulfil their obligations to their participant communities, and to satisfy the expectations of their institutions.

Acknowledgements

Funding: this research was funded by the Te Herenga Waka – Victoria University of Wellington Mātauranga Māori Research Fund – (Project Number:410143).

About the author

Spencer Lilley is an Associate Professor in the School of Information Management at Victoria University of Wellington, New Zealand.

References

Berger, M. (2021). Bibliodiversity at the Centre: Decolonizing Open Access. Development and Change, 52(2), 383-404.

Bryant, T., & Thomas, C. (2024). Black, Indigenous, and Faculty of Color Awareness of Open Access. College & research libraries, 85(1), 7-29.

CONZUL (2022). 5 years of analysing Open Access in Aotearoa: 2017-2021. https://www.universitiesnz.ac.nz/analysisngopenaccess

CONZUL (2021). State of open access in Aotearoa: Year 3. https://www.universitiesnz.ac.nz/stateofopenaccessyear3

Henk,M. (2019). Centring our values: open access for Aotearoa. https://www.tohatoha.org.nz/wp-content/uploads/2019/10/Open-Access-Report-WEB.pdf

Masaka, D. (2018). 'Open Access' and the Fate of Knowledge from Africa: A Theoretical Discussion. The Journal of Negro Education, 87(4), 359-374.

Nature (2024). Publishing options. https://www.nature.com/nature/for-authors/publishing-options#:~:text=The%20difference%20is%20that%20when,8890.00%2F%2412290.00%2F%E2%82%AC10290.00.

Pinfield, S., Wakeling, S., Bawden, D., & Robinson, L. ((2020)). Open Access in Theory and Practice: The Theory-Practice Relationship and Openness. Routledge.

Roa, T., Beggs, J. R., Williams, J., & Moller, H. (2009). New Zealand's Performance Based Research Funding (PBRF) model undermines Maori research. Journal of the Royal Society of New Zealand, 39(4), 233-238.

Saunders, T. (2022). The future is open: Establishing wider open access for research publications in Aotearoa New Zealand. Office of the Prime Minister's Chief Science Advisor _ Kaitohutohu

Mātanga Pūtaiao Matua ki te Pirimia. https://www.pmcsa.ac.nz/2022/07/15/the-future-is-openintern-report-on-open-access-publishing-in-aotearoa/

Smith, A. C., Merz, L., Borden, J. B., Gulick, C. K., Kshirsagar, A. R., & Bruna, E. M. (2021). Assessing the effect of article processing charges on the geographic diversity of authors using Elsevier's 'Mirror Journal' system. Quantitative Science Studies, 2(4), 1123-1143.

Smith, L. (2021). Decolonizing methodologies: research and Indigenous peoples. Zed Books.

Tawhai, V., Pihera, K. & Bruce-Ferguson, P. (2004). Does the PBRF need reshaping? A new Māori educational institution's perspective. Research and Development in Higher Education: Transforming knowledge into wisdom: holistic approaches to teaching 27, 333-340.

White, R. And Remy, M. (2017). University of Otago Open Access Publishing Survey Results

Special version for the Ngāi Tahu Research Consultation Committee https://ourarchive.otago.ac.nz/handle/10523/7333

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Appendix

Glossary of Māori Terms

Hui	Meetings
Iwi	Tribal community
Kaupapa Māori	Māori methodological and theoretical approach to research
Mana	Status (related to a person)
Māori	Indigenous people of New Zealand
Mātauranga Māori	Māori knowledge
Tikanga Māori	Māori customary values
Whanaungatanga	Relationship building

Information Research - Vol. 30 No. iConf (2025)

BSLISE directory: innovating global mapping and organization of LIS schools

Tiago Emmanuel Nunes Braga, Josir Cardoso Gomes, Nathaly Cristine Leite Rocha, Fatih Oguz, and Chris Cunningham

DOI: https://doi.org/10.47989/ir30iConf47272

Abstract

Introduction. The building strong library and information science education (BSLISE) initiative seeks to strengthen the global library and information science (LIS) community by developing an international directory of LIS education programs.

Method. Evolving from the MAP project, the BSLISE directory consolidates and adapts data into a new framework using open-source software. This framework incorporates principles of citizen science and open science, focusing on user-friendly design and integration with existing systems.

Analysis. Data analysis revealed varied regional representation and gaps in accessibility. The directory is structured around organizations, programmers, and users to facilitate learning and knowledge dissemination. Data curation routines and privacy controls ensure data quality and confidentiality.

Results. The directory provides a comprehensive resource for LIS education, incorporating user feedback for continuous improvement. Future developments include integration with RoR and enhanced data categorization. The directory aims to foster international collaboration and incorporate diverse cultural perspectives, benefiting from regular updates.

Conclusion. The BSLISE directory centralizes global LIS education resources, supporting international collaboration and knowledge sharing. Its open-science approach ensures ongoing relevance and adaptability to the global academic community.

Introduction

The building strong library and information science education (BSLISE) is a working group of IFLA that was commissioned in 2018 to develop a set of guidelines for the global library and information science (LIS) community. The BSLISE guidelines serve as a framework for developing LIS education programs, which stakeholders can use for planning, developing, and assessing the quality of LIS education and identifying the necessary knowledge areas for LIS professionals to apply and continue to develop (Chu & Raju, 2018). As part of the guideline development process, a subgroup was formed to design a map project to capture data from global LIS programs and create a visualization of their geographic locations. This map project was populated with data submitted by the global LIS community through a series of surveys released in 2019 and 2020. The call for data submission was shared internationally with the LIS community. The data was cleaned and added to a database used to build the map, which was then introduced and mapped using Visão, an open-source geographic information system (GIS) software hosted by IBICT. The map project's objective was to enable users to focus on a specific geographic area, country, or region and view the available LIS programs within that specified area, as well as other pertinent information. It also aimed to allow employers to quickly find LIS schools and obtain additional information about a candidate's qualifications (Nuria et al., 2022).

However, it was noted that the map primarily addressed geographic access while leaving open additional opportunities for accessing the dataset and improving participation of LIS programs in the project. The main issue was access to textual content, such as detailed descriptions of each program or course, providing deeper insights into their actions and activities. Furthermore, improvements were identified regarding access and discoverability of information, such as the application of thematic filters and individualized access to the page of each school, program, or course. Lastly, the need for a structure that would allow institutions to independently submit and update their data was also considered, promoting greater autonomy and reliability of the information provided. The *BSLISE Directory Project* is an expansion of the Map Project.

Lists of people or organizations which are typically arranged in alphabetical or classified order, providing addresses and other details are generally characterised as directories (Cunha, 2001). For legal entities, this includes the address, names of executives, offered products and services, and similar information. The importance of a digital platform offering a directory structure lies in its ability to centralize essential information, providing a unified point of reference that optimizes data management and access. This mechanism facilitates collaboration among various actors by promoting informational accessibility for all involved. Furthermore, the centralization of information supports interoperability between different systems and platforms, enabling effective compatibility and sharing of critical information. A notable example of a directory is the *Registry of Research Data Repositories* (re3data) (https://www.re3data.org/), which centralizes research data repositories, contributing to discoverability and information preservation.

Given this context, the research focused on organizing an informational structure that supports strengthening the international quality of LIS education and fosters the exchange of information among international schools.

Methodology

To achieve the project's objective, the team adopted different approaches. The informational perspective focused on organizing and structuring the data already gathered as part of the map project and understanding the features necessary to provide the new service. Regarding the computational aspects, the main proposal was to identify an open-source software that would meet the informational requirements. Strategically, the project's results aimed to retain the MAP's most valuable functions, so this was kept in mind throughout the entire process. Finally, the group validated the proposed information architecture.

The first activity was to organize the existing information and understand the available possibilities. The group used the IFLA regional division to prepare a visual presentation of the courses and universities. It was then determined that a directory approach would best fit the informational needs.

The design of the site followed the prototype, assess and review (PAR) method (ROCHA et al., 2024), used in various IBICT projects. This method was applied to identify the most important information structures and was implemented using the spreadsheet already collected by the team members as a starting point. This phase established how the data would be consolidated and which attributes were most important to consider.

Another requirement considered during the project's execution was the use of an open-source software. Using an open-source software is important to foster collaboration and sustainability of the resulting directory in the long term, as it is intended to be operated and maintained by the global LIS community in the future.

IBICT has been working with open science since the 1990s and has experienced different approaches over the decades. One of the projects IBICT has worked on is the *CIVIS citizen science platform*. Once the team involved in the Directory's project analysed the requirements already defined, it was identified that CIVIS could be used as the basis for the new project. This platform is based on the European Commission's citizen science project (https://eu-citizen.science/) and is already multilingual, meaning that other languages could be incorporated in the future. The next steps focused on mapping the specificities that would need to be incorporated into the new system, as well as identifying which existing resources would not be needed in the future platform.

One of the features incorporated during the technology definition process was the online forms. As the intention was to allow users to self-register their schools and courses, this usability was very important. At the same time, it was considered that the submitted data would need to be validated before being made available online. This curation process was also one of the features already incorporated in the CIVIS platform. After all the requirements were verified, the CIVIS platform was defined as the base for the creation of the new BSLISE directory.

Another phase of the research was related to data preparation. The original data was in tabular format and stored in an online Google spreadsheet. In many cases, there was a cardinality problem where an organization contained more than one course, resulting in duplicated educational institution data, which generated inconsistencies and rework. This data had already been validated by the team, but the way the MAP was organized required that this duplicated data be retained. To import the MAP dataset into the Directory database, a Python script was built to define connections between organizations and the courses they offered. The resulting relational database was then checked in three additional aspects: whether the course URL existed, whether there was an image containing the organization's brand available online, and whether the latitude and longitude provided were valid. These verifications were used to ensure a more reliable dataset.

To develop the information architecture of the directory, the initial structure of CIVIS was used as a foundation, and the information fields were adapted to meet the specific needs of the LIS courses data. From the available fields, the following were selected:

- **Programs:** name, URL, description, qualification, theme, geographic scope, countries, organizations, contact, email, phone, profile image
- **Organizations or networks:** name, URL, description, organization type, contact, email, organization logo, location
- **Users:** name, title, email, country, biography, areas of interest, profile picture

This set of informational fields allows the data recorded in the directory to be used not only for visualization purposes but also for creating information refinement filters and establishing relationships between programs and organizations.

Finally, the proposed architecture was validated first by professionals involved in the MAP project and then by the BSLISE group. Comments and contributions were aggregated and incorporated if viable.

Results and discussions

The first result achieved was the understanding that some regions were more represented in the MAP than others. This was possible due to a visual organization of all courses submitted to the MAP project. The tool used was a sunburst graphic structured based on the six IFLA regions: Asia-Oceania, Europe, Latin America and the Caribbean, Middle East and North Africa, North America, and Sub-Saharan Africa. The result is presented in Figure 1 below.

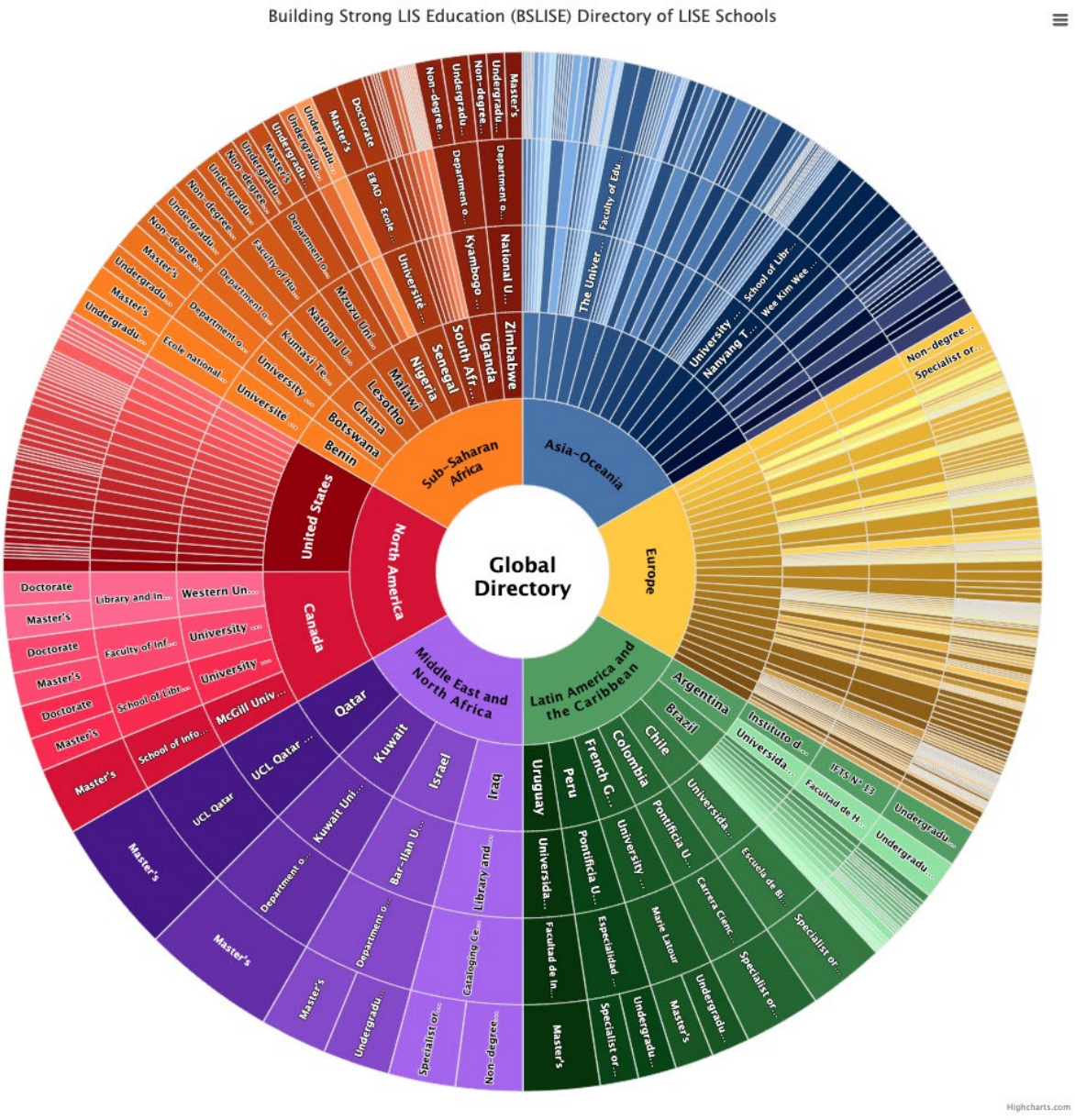


Figure 1. Sunburst graphic

The sunburst graphic highlighted that some regions were denser than others in terms of registered schools and courses, such as Europe and North America, by the other hand, Sub-Saharan Africa, and the Middle East show less representation, highlighting potential gaps in data collection. The tool used

effectively visualized density by region but could not identify duplicate entries. One aspect that the team working on the project understood was that the IFLA regions did not need to be a requirement for the project, as there would only be a visualisation filter generated automatically. Hence, this characteristic would not be incorporated into the first version of the repository.

Data modeling leveraged IBICT's experience in setting up CRIS systems (Segundo et al., 2023) that use standardized international ontologies. The process integrates the requirements suggested by the BSLISE working group with the state of the art in modeling the organization and course entities. The final model is described in Figure 2.

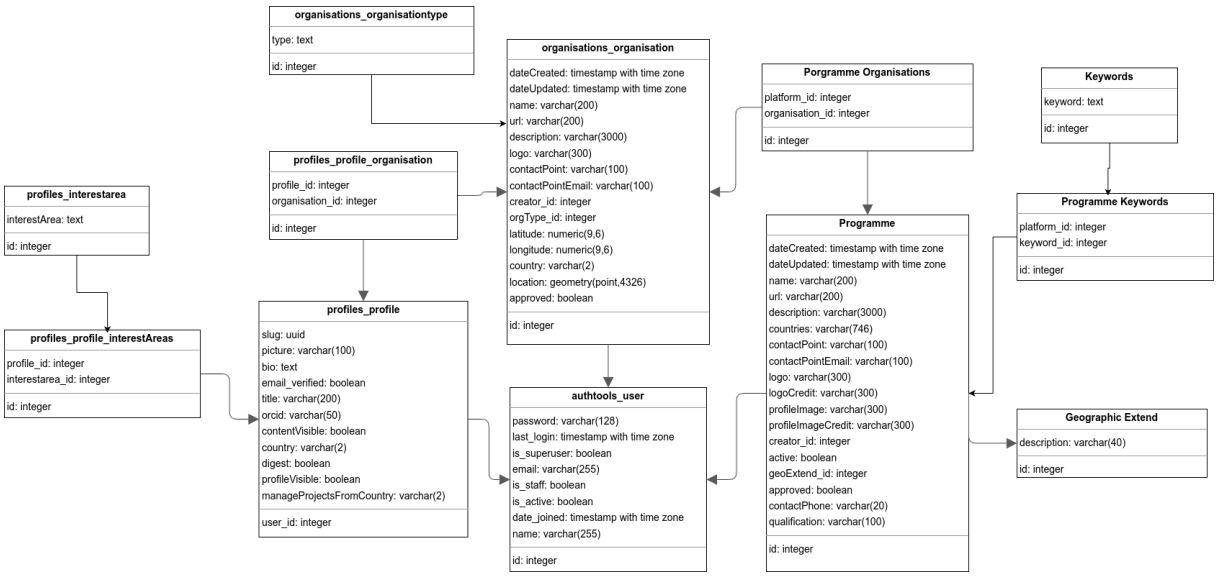


Figure 2. Main model

The visual proposal for the new directory considered that the platform should be easy to use and understand. The MAP project required the creation of manuals, which took some months to complete, and this was not the intention of the directory. Based on the three entities—organizations, programs, and users—a very clean information architecture was proposed, as shown below.

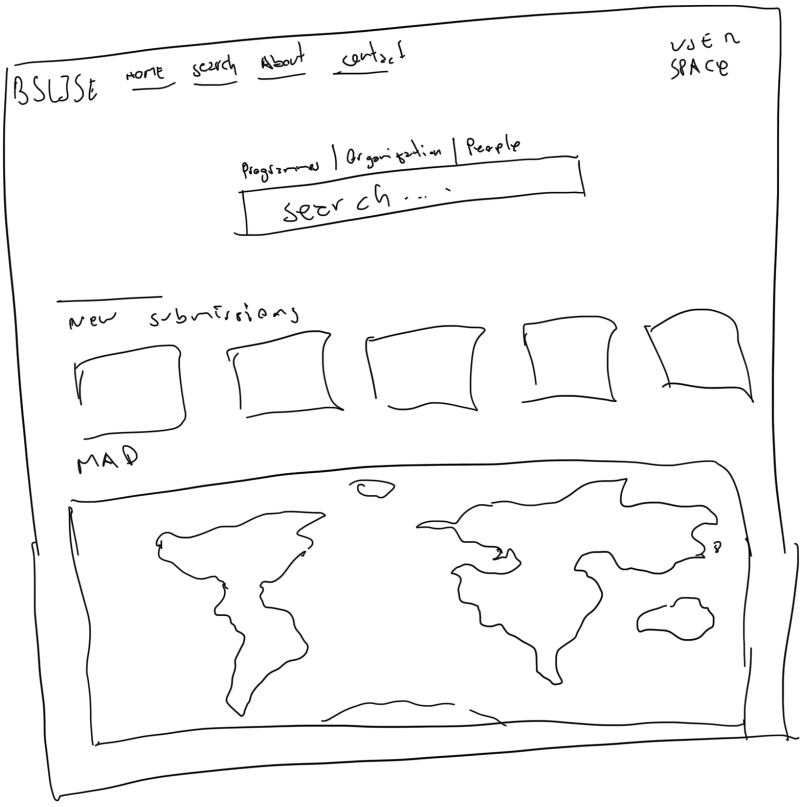


Figure 3. Prototype of information architecture

After applying the proposed data methods, the data obtained in the MAP phase was consolidated based on the directory proposal. More than 170 courses were registered and mapped according to the organizations running each course. Different metadata were mapped, and the organizations were categorized mainly into academic, governmental, and non-governmental.

The information architecture, supported by the pillars of '*organization*, *courses*, *and users*', has been strategic, as these three components encompass the main dimensions involved in learning and knowledge dissemination. Organizations provide an institutional reference point, ensuring credibility and a variety of options. Courses, in turn, are the informational core, representing the available educational offerings with varying levels, durations, and formats. Finally, users include both students and professionals interested in qualifications or updates, allowing the directory to be tailored to meet the needs and preferences of this audience. This triad ensures an integrative approach, facilitating efficient access to information in a way that aligns with the needs of the public.

Routines have also been implemented to curate the data entered by external users. The process runs as follows: the user registers their data on the platform, and the system checks that the email is valid and sends a notification to the registered email. There is also a CAPTCHA check to prevent attacks. Based on this automated validation, curators verify the data via ORCID or other sources and approve the user's registration.

Strict control is maintained over user privacy, allowing users to define at any time which data will be public and which will only be available to the platform's managers. Once approved, users can register organizations and courses, which will also go through the same curation process. The entire curation process is conducted directly on the platform, without the need for external controls.

The development of the BSLISE directory was supported by two key concepts: citizen science and open science. Regarding citizen science, the design of the directory highlights the use of two principles described by the European Citizen Science Association (2015): 1. citizen science projects actively involve

citizens in scientific activities, generating new knowledge and understanding and 7. data and metadata resulting from citizen science projects are made publicly available and, whenever possible, published in an open-access format. Both principles are closely aligned with the concept of open science. While the proposed information flow ensures that anyone aiming to contribute to the project can do so, it also guarantees that the submitted data adheres to the quality standards proposed by the BSLISE group. Additionally, every aspect of the directory is open, from the software used to store the data. All data related to submissions marked as public can be downloaded and integrated with other software or platforms.

The final version of the directory is shown below and will be accessible in the future using the link https://directory.bslise.org.

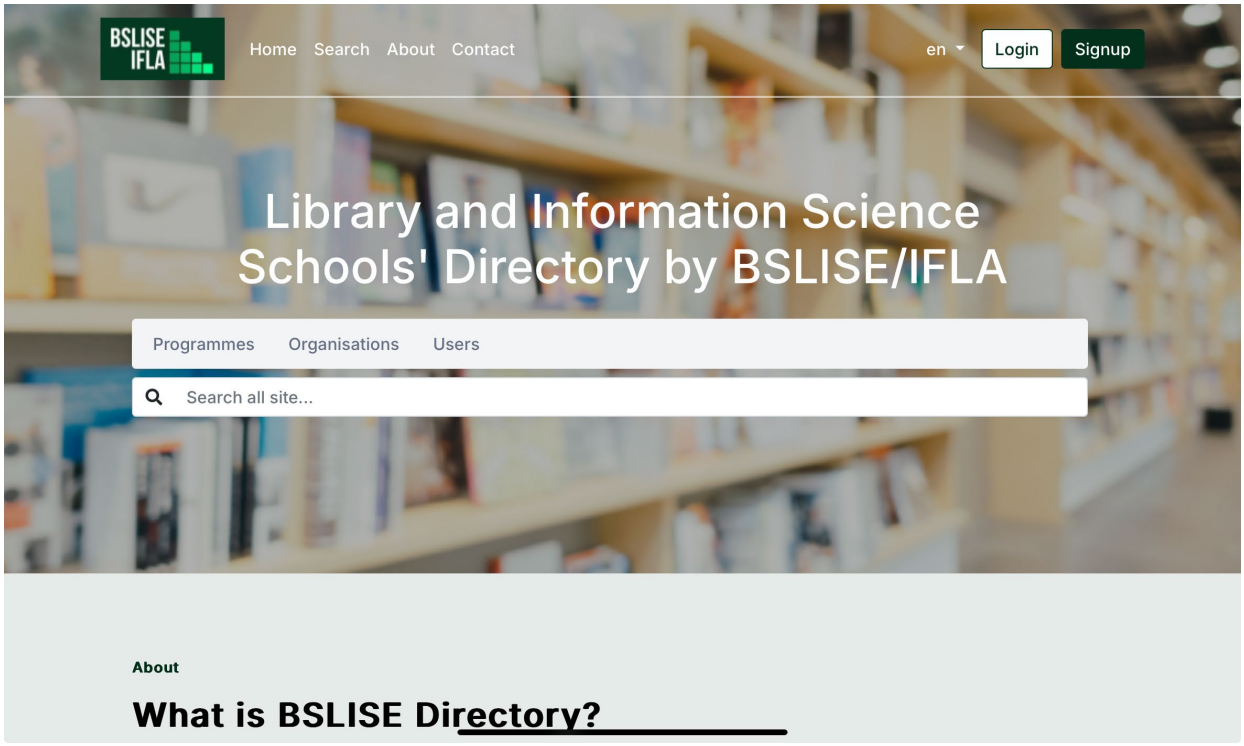


Figure 4. Screenshot of the directory's homepage

Finally, the members of the BSLISE group were able to access and provide input to contribute to the directory project. The group discussed the directory's characteristics as the project progressed.

Conclusion

The implementation of the BSLISE directory is a work in progress. New routines and adjustments continue to be made as the academic community becomes aware of the initiative and suggests improvements and new features. Also under construction are routines for integration with RoR, re-categorization of organization types to simplify data entry, and language customization for the emails sent to registered users in the directory.

Once the collaboration network is fully operational, there are plans to extend the system so that users can add curricula taught by the courses and educational resources useful to the academic community. For the LIS community, the directory can serve as a starting point for deepening international collaboration. The citizen science perspective allows the project to incorporate different cultural approaches, which could enhance the directory's adoption by the global LIS community. If successful, the directory will benefit from frequent data updates and validations.

It is important to highlight that the Directory project maintained the consistency of the data previously submitted to the MAP project. This was achieved because both projects were discussed and developed within the scope of the BSLISE group and followed the requirements established by the professionals involved in this initiative.

Acknowledgements

We sincerely thank the Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) and the Building Strong Library and Information Science Education (BSLISE) for supporting and contributing to this research. Their insights and resources were invaluable to the development of this work

About the authors

Tiago Emmanuel Nunes Braga is the Director of the Brazilian Institute of Information in Science and Technology (Ibict) and a professor in the Postgraduate Program in Information Science at Ibict. With a Ph.D. in Information Science from the University of Brasília (UnB), his expertise spans data science, digital preservation, citizen science, and information for sustainability. Tiago can be contacted at tiagobraga@ibict.br

Josir Cardoso Gomes holds a Ph.D. in Information Science from the PPGCI/IBICT/UFRJ, a Master's in Administration from IBMEC/RJ, and a degree in Data Processing Technology. He is the Vice-Coordinator of the Laboratory in Network of Digital Humanities (Larhud) and a professor at the School of Communication, Media, and Information at FGV/RJ. Josir can be contacted at josirgomes@ibict.br

Nathaly Cristine Leite Rocha holds a master's degree in Information Science from the University of Brasília (UnB) and is a researcher at the Brazilian Institute of Information in Science and Technology (Ibict). Her research interests include data visualization, georeferenced information systems, and digital document indexing. Nathaly can be contacted at nathalyrocha@ibict.br

Fatih Oguz is an Associate Professor in the Department of Information, Library, and Research Sciences at the University of North Carolina at Greensboro. He earned his Ph.D. in Information Science from the University of North Texas, an M.B.A. from Yeditepe University, and a B.S. in Chemistry from Boğaziçi University. His research interests include information access, digital libraries, and online education. Fatih can be contacted at f_oguz@uncg.edu

Chris Cunningham is an Assistant Professor at North Carolina Central University (NCCU), specializing in Library and Information Sciences. He teaches courses such as Management and Leadership of Information Organizations and Libraries. He earned his Ph.D. from the University of North Carolina at Greensboro. Chris can be contacted at chris.cunningham.lis@gmail.com

References

Cunha, M. B. (2001). Para saber mais: Fontes de informação em ciência e tecnologia. Briquet de Lemos/Livros. http://repositorio.unb.br/handle/10482/15121

Chu, Clara M & RAJU, Jayarani. (2018). Building strong LIS education: a call to global and local action: an IFLA BSLISE Working Group white paper. [S. l.]: IFLA.

ECSA (European Citizen Science Association). 2015. Ten Principles of Citizen Science. Berlin. http://doi.org/10.17605/OSF.IO/XPR2N

Puig, N. B., Braga, T. E. N., Oguz, F., & Cunningham, C. (2021). An international approach on the GIS mapping of library and information science courses. Tendências da Pesquisa Brasileira em Ciência da Informação, 14.

Rocha, N., Alves, L. D. A., & Braga, T. E. N. (2023). Metodologia colaborativa para levantamento de necessidades informacionais em projetos de pesquisa. Anais VI WIDAT, 1–13. https://doi.org/10.22477/vi.widat.76

Segundo, W., Dias, T. M., Moreira, T., Pinto, A. L., Silva, V., Gomes, J., Quoniam, L., Matas, L., Dias, A., & Schneider, J. (2022). A strategy for collection, integration, and processing of scientific data in BrCris context. In T. M. Rodrigues Dias (Ed.), Advanced Notes in Information Science (Vol. 2). Pro-Metrics. https://doi.org/10.47909/anis.978-9916-9760-3-6.117

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Towards nuanced and critical data analysis: an exploratory investigation into complexities of Goodreads reviews for children's books

Yuerong Hu and Chen Ling

DOI: https://doi.org/10.47989/ir30iConf47269

Abstract

Introduction. With the advance of artificial intelligence and big data, digital humanities have been empowered with new research affordances. However, computational modelling on cultural datasets without a critical examination of the datasets risks decontextualized datafication and misinterpretation, which does not necessarily advance our knowledge into humanities and social sciences questions. This paper presents an exploratory investigation into 37,408 Goodreads reviews on children's books to showcase the complexities of cultural datasets online, to inform contextualized and critical cultural data analysis.

Method. We conducted topic modelling and qualitative analysis of content on the reviews retrieved by regular expressions and visualized our findings.

Analysis. Our analysis empirically illuminates (1) participation in reviewing children's literature online, and (2) presentation of topics and opinions in the reviews.

Results. We found that although various family members across generations participate in reading children's books and reviewing them online, female members play particularly active roles, and children's opinions are underrepresented.

Conclusions. In addition to gaining insights into reviews on children's books, this study also demonstrates the importance of exploratory data analysis in gaining a more nuanced understanding of cultural datasets online, which is essential for advancing insightful and critical data analysis.

Introduction

In the last decade, artificial intelligence has become increasingly impactful in both socio-cultural life and humanities research, from generating poetry that competes against '*the work of professional writers*'(Köbis and Mossink, 2021) to deciphering ancient scrolls (Marchant, 2024). However, tensions persist between humanities methodologies and AI technologies (Chun and Elkins 2023). One key question is AI applications' capabilities to depict the complexities and nuances of humanities data and questions (Chun and Elkins 2023; Windhager and Mayr, 2024). This is a particularly pressing problem for digital humanities research on social media data '*produced within the fabric of sociomaterial*' (Wu, 2023). Due to factors ranging from the platform owners' restricted data access to heterogeneous user behaviour, rarely can scholars comprehensively evaluate the limitations, biases, and representativeness of the datasets available for research. This lack of understanding of the datasets remains a bottleneck for critical usage of large humanities datasets, despite advances in AI (Nguyen et al. 2020).

To address this problem, digital humanities and information sciences scholars have proposed multiple theories and methods, such as data archaeology (Lee, 2021), predictive modelling (Underwood 2019), and speculative bibliography (Cordell, 2020). It is increasingly acknowledged that a nuanced understanding of the data as it is (e.g., how the data has been created, curated, and used within social, cultural, and technical systems) is key to computational modelling that adequately reflects the dataset's complexities and contexts.

Digital humanities research on online book review studies provides a modelling research case. Although advanced AI models have been applied to reviews across various platforms and languages, answers to some research questions remain unclear (Herrmann et al., 2024). For example, are all reviewers merely book readers (Landow, 2006)? To what extent are reviews incentivized (Hu et al., 2023), copied (David and Pinch, 2006), or generated by AI? Why do some reviewers rate their favourite books with 1 star and less favoured ones with 5 stars (citation anonymized for blind review)? Without addressing these questions, AI applications can hardly make reliable and insightful interpretations of millions of book reviews available.

Among various book reviews, those for children's books are particularly challenging to interpret computationally as they tend to involve multiple stakeholders and perspectives. For example, many self-identified parent reviewers provide opinions of their own and those of their children in the same review; therefore, opinions and sentiments presented in the same review might diverge drastically. Yet only one numeric rating is associated with each review. Such features pose practical questions to data analysis: to what extent and how are children readers' opinions represented by online book reviews, and how can researchers identify them? When multiple opinions are presented in one review, how can the computer disentangle the discourses or enrich the single-valued ratings?

It is essential to explore these questions for more advanced computational data analysis, which advances humanities understanding of children's literature and yields practical implications. For decades, a number of heated discussions on children's literature remains theoretical or anecdotal, partially due to the lack of reading records (Hu, et al. 2023): Who is in a position to judge children's literature, adults, children, or both (Abate, 2020)? What are the perceived differences between literature for children and young (Gubar, 2011; Trites, 2001)? Unprecedented online book reviews enable scholars to empirically understand the reception of children's literature from different perspectives and with varied granularities, answering existing research questions while raising new ones.

Given such backgrounds, we conducted exploratory data analysis on 37,408 reviews on 17,407 children's books collected from the social reading platform Goodreads. This is designed as an exemplar case study of inspecting characteristics of social media datasets for advancing cultural

data analytics. We leveraged both computational models and qualitative analysis of content to explore three questions: Who engages with reading and reviewing children's literature? How are topics and opinions presented in the reviews? Based on the first two questions, what are the practical implications for AI-infused research on children literature reviewing and datasets alike? Specifically, our research focuses on reviews that explicitly mention family members, as reading and reviewing children's books within families is a prominent setting identified in our dataset. In the remainder of this paper, first, we introduce the research dataset and method. Then we present our data analysis and outcomes. Finally, we summarise our findings and their limitations, along with proposals for future work.

Dataset and methods

We leveraged existing research datasets called Goodreads book graph datasets (Wan et al. 2019; Wan and McAuley 2018). The data was scraped from Goodreads, a social reviewing platform based in the U.S. and one of the most prominent data sources for digital humanities research on online book reviewing (Walsh and Antoniak 2021). Our analysis focused on its subset of reviews for children's books. We pre-processed the reviews by dropping the reviews that are not in English or too short to be semantically meaningful (with less than 50 English characters). This process returned 297,222 reviews. Then, we developed a small dictionary for computationally identifying the reviews that mentioned family members. Based on our preliminary data annotation and testing, the regular expression of 'my' followed by a singular or plural form of a noun for a family member effectively and efficiently retrieves reviews of our research interest. Please see Appendix 1 for the family members included in the dictionary. We include 'friends' as an exceptional 'family member' because our pilot research shows that friends are often referred to with similar intimacy and significance as one's family members. With this dictionary and regular expression matching, we retrieved 37,408 reviews that mention family members. It is to be noted that neither our dictionary nor the regular expression-based method can exhaustively identify all reviews mentioning family members, as our goal was to identify sufficient instead of all reviews in this category for exploratory analysis.

We applied two complementary analyses to the reviews. First, we conducted a computational overview of the review regarding (1) mentions of family members and (2) topics that emerge from the reviews. Second, we conducted a qualitative analysis of content (Wildemuth, 2017) on a stratified sample of 400 reviews. For topic modelling, we used the topic coherence measure implemented in Gensim (Rehurek and Sojka, 2011) to optimise the number of topics: using LDA modelling and searching across a range from 10 to 90 topics, the topic coherence optimised at 40. Therefore, we set the number of topics to 40. Please see the Appendix 2 for details of the topics. For content analysis, we randomly selected an even number of reviews associated with each of the family members appearing in the reviews. After deduplication, we got a sample of 400 reviews which we manually annotated and analysed.

Our research proposal was reviewed by the Institutional Review Board (IRB) of our current institution. A decision was made that the application for non-human subjects' research or IRB application was not required. Nevertheless, our research complies with the guidelines proposed by latest research on the legal and ethical use of online book reviews in research (Hu, Layne-Worthey, et al. 2023). For instance, reviews quoted in this paper were all paraphrased and further anonymized to minimise the risks that might be caused to reviewers (e.g., reidentification).

Analysis and findings

Family participation in reading and reviewing children's books

First, to understand who participates in reading and reviewing children's literature in a family, we checked the occurrences of each cluster of family members in the 37,408 book reviews, as is summarised in Table 1. We grouped the 66 named entities of family members that appear in the

reviews into 19 primary groups for presentation purposes. We can tell from Table 1 that the group of words for '*my child*' (including '*my child*', '*my daughter*', '*my son*', and many others), appear 36,383 times in the 37,408 book reviews. Figure 1, the sunburst plot, visualises the distributions of the occurrences in Table 1 in a hierarchical manner: the inner ring represents numbers of occurrence in groups; the outer ring shows how the groups are broken into more detailed counts, with numbers after the words indicating their total occurrences. According to the counts, a variety of family members were mentioned 56, 625 times in total. The most frequently mentioned members are the reviewers' children (64%, n=36, 383), siblings (6%, n=3,393), and friends (5%, n=2,793). The reviewers' parents also appear many times, including mother (5%, n=2,741), parent (3%, n=1,928), and father (2%, n=1,249). Other frequently mentioned family members include reviewers' grandchild (3%, n=1,756), grandparent (2%, n=1,023), and family (as a whole) (2%, n=958).

In addition, the weight of female family members stands up (based on conventional constructions of family members' gender identities). When referring to family members senior to the reviewers, female family members are more likely to be mentioned. Despite wording differences, the occurrences of a reviewer's mother vs. father are 2,741 vs. 1,249, the occurrences of grandmother vs. grandfather are 485 vs.290, and the occurrences of aunt vs. uncle are 137 vs. 74. In these three cases, mentions of female members approximately doubled mentions of their male counterparts. The unbalance is more apparent when reviewers mention their partners/spouses mentions of husband (n=1,204) triple mentions of wife (n=367) and mentions of boyfriend (n=125) quadruple girlfriend (28). However, mentions of younger generations in the family are more balanced gender-wise. For instance, reviewers' sons are mentioned 6,743 times, while their daughters are mentioned 7,212 times; grandsons are mentioned 496 times, and granddaughters are mentioned 488 times. One possible explanation for these findings is that female readers/reviewers have been more actively involved in reading children's literature at home and reviewing books online: When female readers review books online, they may mention their spouses or partners who are often male, as well as their children of various genders; on the other hand, when non-female family members write book reviews, they often mention their female family members who play a significant role in reading at home.

group	entity	count	total	group	entity	count	total
	child	6,708			grandmother	309	
	daughter	6,632			grandpa	171	
	son	6,402			grandma	160	
	kid	5,521			grandparent	131	
	kids	4,864			grandparents	117	
	children	3,474			grandfather	115	
	boys	842			grandmothers	16	
	girls	681		grandparent	grandfathers	4	1,023
	daughters	580			niece	1,051	
	sons	341		niece	nieces	263	1,314
	baby	280			nephew	546	
	babies	32		nephew	nephews	144	690
child	twins	26	36,383		mother-in-law	32	
	friend	1,782		parent-in-law	father-in-law	13	45
friend	friends	1,011	2,793		sister-in-law	30	
	mom	1,395		sibling-in-law	brother-in-law	12	32
mother	mother	1,346	2,741		aunt	130	
	dad	722		aunt	aunts	7	137
father	father	527	1,249		uncle	68	
	parent	983		uncle	uncles	6	74
parent	parents	945	1,928		sister	1,117	
family	family	958	958		brother	690	
husband	husband	1,204	1,204		younger	638	
wife	wife	367	367		older	428	
	granddaughter	439			sisters	115	
	grandson	433			sibling	115	
	grandchild	390			siblings	110	
	grandchildren	382			brothers	89	
	grandsons	63			twin	62	
grandchild	granddaughters	49	1,756	sibling	elder	29	3,393
	cousin	233			boyfriend	125	
cousin	cousins	61	294		partner	54	
					girlfriend	28	
				partner	spouse	27	234

Table 1. Occurrences of the family members being mentioned in the reviews

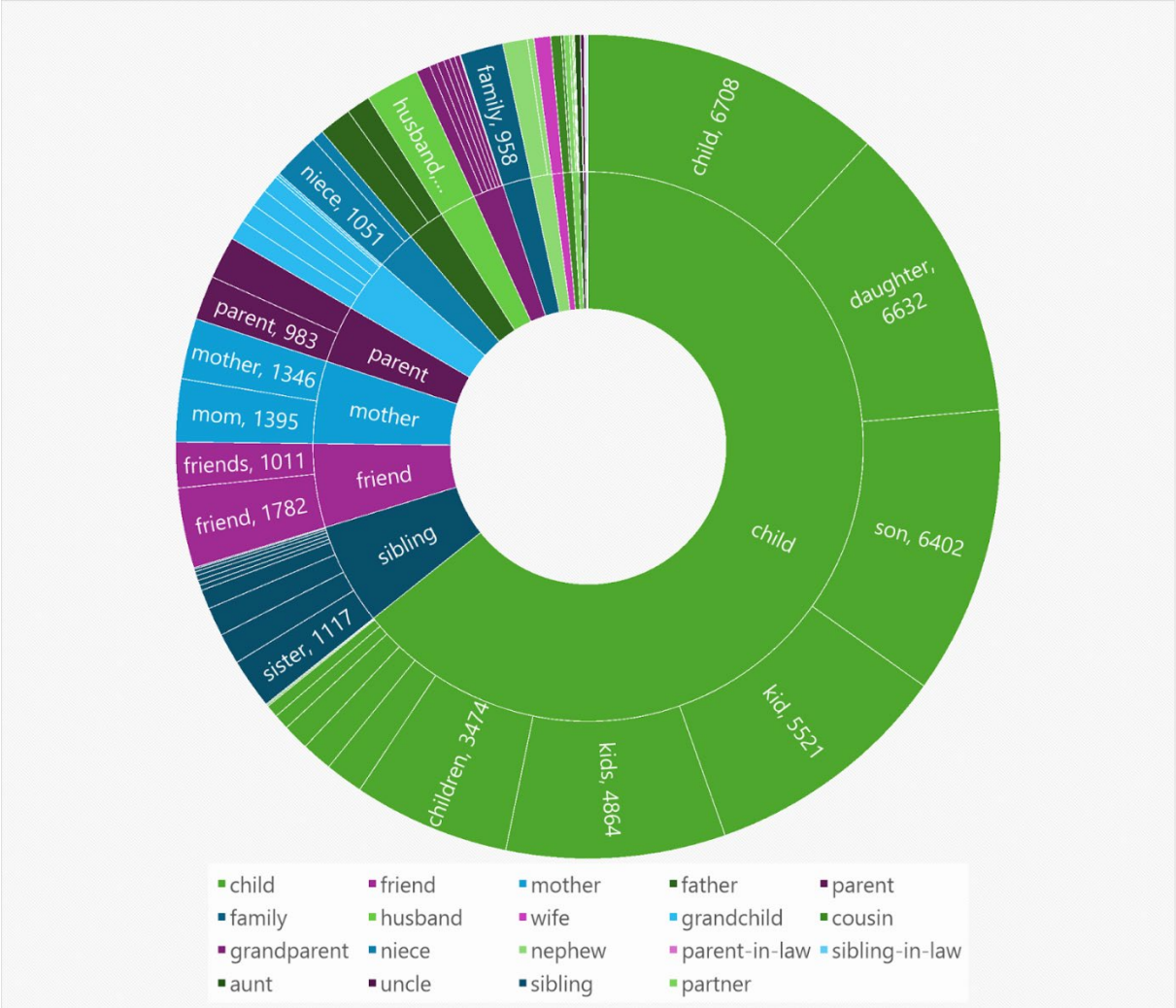


Figure 1. A sunburnt plot depicting distributions of occurrences of family members

Furthermore, we analysed the co-occurrence of different groups of family members in the same reviews. Figure 2 provides an overview: the number in each coloured block indicates the number of times that two different groups of family members appear in the same review; the corresponding textual variables on this coloured block's x-axis (the horizontal line) and y-axis (the vertical line) tell which two groups of members they are. The colour of the block represents the relative frequency of the paired codes' co-occurrence: the darker colours indicate that the paired codes were brought up together more frequently, while the lighter colours suggest fewer co-occurrences. For instance, we can tell from the darkest blue block that reviewers' siblings and children were mentioned simultaneously in 482 reviews, which indicates a stronger proximity or correlation between the two groups. Reviewers' children also co-occur with the reviewers' siblings' mothers (n=452), husbands (n=340), and friends (n=208). In addition, reviewers' mothers and siblings frequently appear in the same reviews. Our close reading reveal that this often happens when reviewers write about their childhood memories of story time with their mothers and siblings, which echoes the hypothesis that female members play a significant role in reading children's books at home.

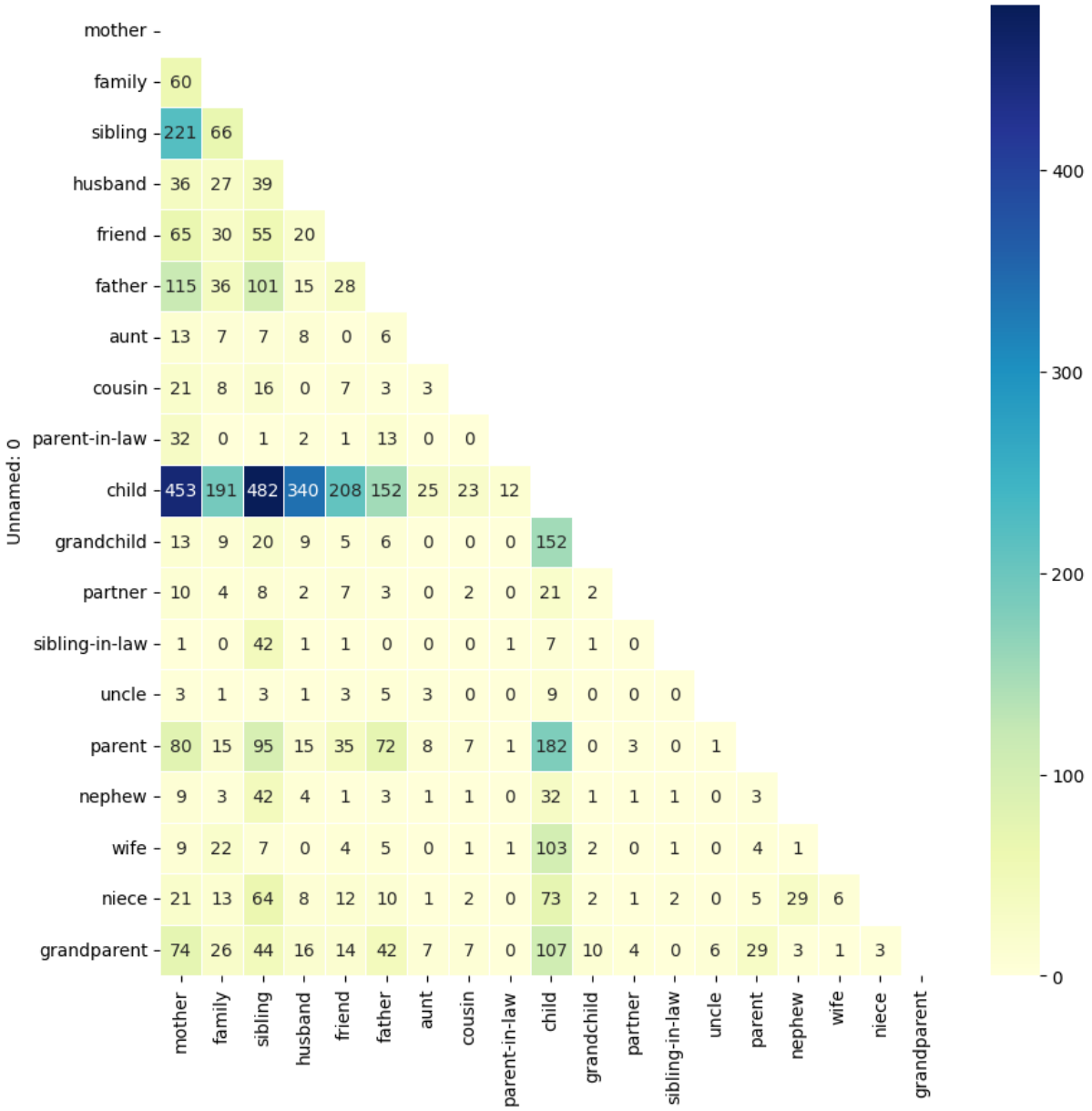


Figure 2. Co-occurrences of family members mentioned in the reviews

Topics and opinions presented in the reviews

First, our topic modeling on the 37,408 reviews provides an overview of topics discussed. In addition to family members, topical words center on book genres and features (e.g., illustration, version, and graphic novel), sentiment and emotions (e.g., amazing, humorous, and thrill), education and schooling (e.g., teacher, school, and study), as well as reading behaviour and scenarios (e.g., reread, bedtime, and listen). For instance, words associated with fairytales and fantasy stories are also commonly seen (e.g., magic, mystery, and enchant), which aligns with genre features of children's literature. Comparing mentions of family members with the topics, we find that reviews involving senior family members, particularly grandparents, are more likely to involve nostalgia emotions, childhood memories, and retrospective reflections. It is also noted that many topic words are conventionally associated with female characteristics, such as actress, witch, ballet dancer, and princess, which might reflect gender stereotypes in children's literature.

Second, we applied close reading and qualitative analysis to the sample of 400 reviews. 70 reviews explicitly mentioned children's reactions to the book being reviewed, while none of the reviewers self-identified themselves as children. Most of the reviews were created (1) to share personal opinions about a children's book as a child supervisor and/or care provider; (2) to express retrospective opinions about a book they had read in their childhood; and (3) to express their thoughts about a book they recently got introduced to or re-read.

These suggest that the reviews might be dominated by adults' opinions. Specifically, according to the reviews written by self-identified parents, parents tend to make decisions for their children, and sometimes even impose their preferences on their children. For instance, one parent wrote:

> *if I cannot bear with a book, I'm not going to read it to my kids by giving it away or hiding it somewhere they won't find or request the book.*

Children's opinions might be underrepresented in numeric book ratings as well. For instance, a parent gave a 3-star rating to a book and commented '*I read this book to my girls who are 5 and 11. They both loved it; but I was bored*'. While acknowledging their children's love for the book, this review did not give it 4-star or 5-star ratings, which are generally regarded as positive. A few parents further explained why they failed to relate to their children's favor for a book, such as '*my children love this book, probably because it contains the word fart, but I do not think this is one of my favourite books on the shelves*'.

Another parent commented:

> *my children love this book because I cannot help crying when reading it aloud to them, then they point at me and laugh, when their father weeps, trying to get the words out without turning into a sobbing mess. sigh. I do love my damn kids.*

Such disagreements and divergences of opinions across reviewers and their children also exist between reviewers and their more senior family members. For instance, a reviewer wrote '*I will never get why my grandmom was so fond of this book. Who needs an upsetting story about cruelty to animals? I am not*'.

Nevertheless, there are also cases where opinions about books converge across generations and books get to be passed on in a family, as the review below shows:

> *my grandma used to read this book to me when I was a kid every time I stayed overnight at her place. I still vividly remember how I sat on her lap in a recliner and listened to her voice. I can still hear it. One day I will read this book to my own kids.*

Out of the 70 reviews, only two offer perspectives from multiple children, as shown in Figure 3. The first review includes the opinions of two child readers but does not reference the adult reviewer's perspective. The second review presents ratings and feedback from each family member, including two adults and three children. For such reviews that feature multiple sets of opinions and ratings from different individuals, the standard approach of treating one review as representative of a single reviewer no longer applies. A more nuanced data analysis, accommodating varying granularities of opinions, is needed for a better interpretation of the data.

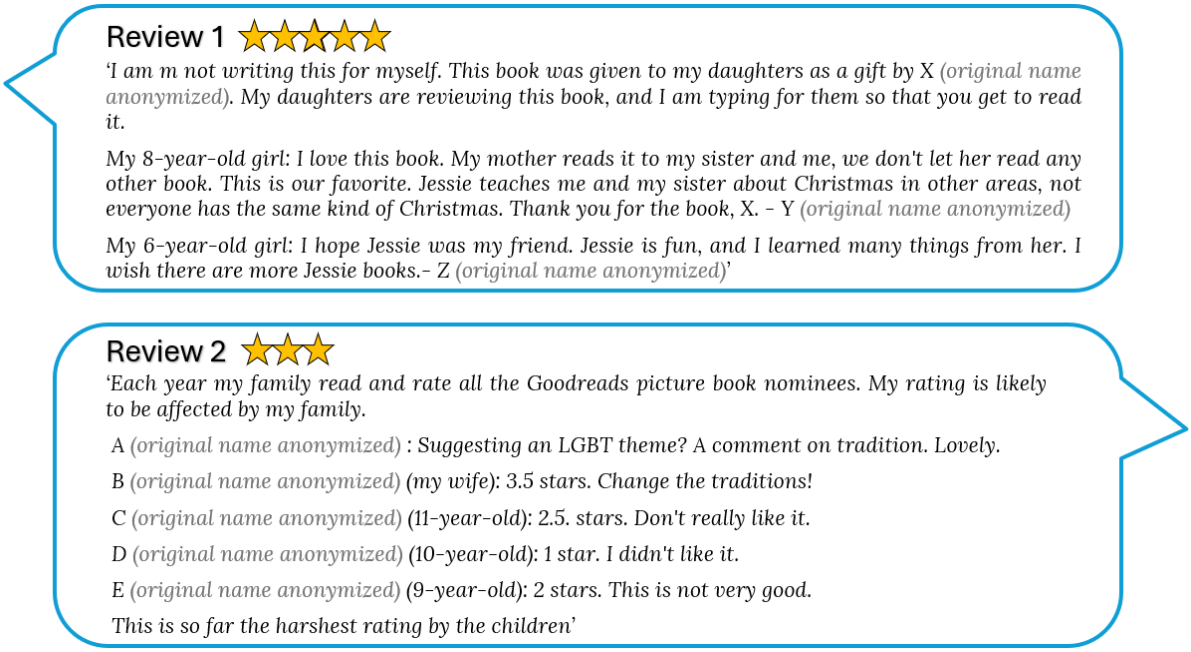


Figure 3. Two reviews that present multiple children's opinions

Conclusions and discussions

This case study of 37,408 Goodreads reviews on children's books has three primary findings. First, a variety of family members were mentioned in online book reviews on children's books, which suggests their broad participation in reading children's literature. Parents and female family members played especially prominent roles in reading such books in households and reviewing them online. Second, while various topics were discussed in the reviews, the reviews were predominantly adults' voices. Children's feedback on books appears to be underrepresented in both review texts and numeric ratings. Third, findings of our study, such as the gender-specific topic words and some parents' irrelevance to children's love for certain books, speak to problems identified by critical scholarship on children's literature, such as sexism and gender issues in children's literature (Varga-Dobai 2013), paternalism (Gubar, 2020) and censorship (Jenkins, 2010; MacLeod, 1983) in children's reading activities. Such connections demonstrate the promise of enriching critical scholarship on children's literature with online book reviews and empirical methods. Questions worthy of future investigation also emerge from our preliminary study, such as (1) how cross-generation readers of the same family negotiate with each other, given divergent opinions about the same books; and (2) how female family members' participation might have shaped online book reviewing.

As an exploratory analysis of a specific dataset, these preliminary findings are subject to further examination and might not apply to other datasets. Therefore, we intend further examine the prominent roles played by female family members through hypothesis testing and analysis on other datasets. Nevertheless, existing findings demonstrate the complexities of online reviews of children's books and have practical implications for cultural data analysis. For instance, this case study demonstrates how mixed-method analysis contributes to a more comprehensive understanding of online cultural datasets; such understanding should help researchers identify suspicious data instances such as fake reviews and AI-generated content, so as to improve data quality in scholarly research. In the future, we plan to deepen our analysis by conducting more qualitative investigations and testing our findings on other datasets. We are currently working on surveys on online book reviews and interviews on reading children literature at home, to advance our understanding of the dissemination and evaluation of children's literature.

Acknowledgements

We thank the Goodreads book reviewers who shared their ratings and reviews. This research would be impossible without their contributions. We also want to thank anonymous reviewers of iConference 2025 for their critical feedback.

About the authors

Yuerong Hu is an Assistant Professor at the Department of Information and Library Science of Luddy school of Informatics, Computing, and Engineering, Indiana University Bloomington, 700 N. Woodlawn Ave, Bloomington, Indiana, 47408, USA. She received her Ph.D. from University of Illinois Urbana-Champaign. Her research combines digital humanities, cultural analytics, and critical data studies to advance contextualized analysis of cultural datasets. She can be contacted at yuerhu@iu.edu.

Chen (Cici) Ling is an Assistant Professor at Informatics Department of Luddy School of Informatics, Computing, and Engineering at Indiana University Bloomington, 700 N. Woodlawn Ave, Bloomington, Indiana, 47408, USA. She received her PhD from Boston University. Her research spans security, privacy, psychology, and computational social science, with a particular focus on social media and niche communities, aiming to promote accessibility and inclusivity in digital applications. She can be contacted at ccling@iu.edu.

References

Bannon, L. (2011). Reimagining HCI: toward a more human-centered perspective. Interactions, 18(4), 50–57. https://doi.org/10.1145/1978822.1978833

Abate, M. (2020). No Kids Allowed: Children's Literature for Adults. Baltimore: Johns Hopkins University Press. https://muse.jhu.edu/pub/1/monograph/book/78274.

Chun, J, & Elkins, K. (2023). The Crisis of Artificial Intelligence: A New Digital Humanities Curriculum for Human-Centred AI. International Journal of Humanities and Arts Computing 17 (2): 147–67. https://doi.org/10.3366/ijhac.2023.0310.

Cordell, R. (2020). Speculative Bibliography. Anglia 138 (3): 519–31. https://doi.org/10.1515/ang-2020-0041.

David, S, & Pinch, T. 2006. "Six Degrees of Reputation: The Use and Abuse of Online Review and Recommendation Systems." First Monday 11 (3).

Gubar, M. (2020). Toothless Pedagogy? Problematizing Paternalism in Children's Literature and Childhood Studies. Children's Literature 48 (1): 153–97. https://doi.org/10.1353/chl.2020.0007.

Herrmann, B., Pianzola, F., van Zundert, J. J., Hu, Y., Nielbo, K., Koolen, M., ... & Boucher, M. C. Unsiloing and Unsilencing Online Reader Reviews - Mapping Reception of Fiction Novels Across Cultures and Languages. In Proceedings of the 11th Edition of the DH Benelux Conference. Belgium. https://2024.dhbenelux.org/wp-content/uploads/2024/05/DHB24_paper_Herrmann_et-al.pdf.

Hu, Y., Layne-Worthey, G., Martaus, A., Downie, J.S., Diesner, J. (2023). Research with User-Generated Book Review Data: Legal and Ethical Pitfalls and Contextualized Mitigations. In: Sserwanga, I., et al. Information for a Better World: Normality, Virtuality, Physicality, Inclusivity. iConference 2023. Lecture Notes in Computer Science, vol 13971. Springer, Cham. https://doi.org/10.1007/978-3-031-28035-1_13

Hu, Y., LeBlanc, Z., Diesner, J., Underwood, T., Layne-Worthey, G., & Downie, J. S. (2024). Complexities of leveraging user-generated book reviews for scholarly research: transiency, power dynamics, and cultural dependency. International Journal on Digital Libraries, 25(2), 317-340. https://doi.org/10.1007/s00799-023-00376-z.

Jenkins, C. A. (2011). Censorship: Book Challenges, Challenging Books, and Young Readers. In Handbook of Research on Children's and Young Adult Literature (pp. 443-454). Routledge. Kidd, Kenneth B. 2007. “Prizing Children’s Literature: The Case of Newbery Gold.” Children’s Literature 35 (1): 166–90.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in human behavior, 114, 106553. https://doi.org/10.1016/j.chb.2020.106553.

Landow, G. P. (2006). Hypertext 3.0: Critical theory and new media in an era of globalization. The Johns Hopkins University Press.

Lee, B. (2021). Compounded mediation: A data archaeology of the newspaper navigator dataset. Digital Humanities Quarterly, 15(4).

MacLeod, A. S. (1983). Censorship and children's literature. The library quarterly, 53(1), 26-38.

Marchant, J. (2024). First passages of rolled-up Herculaneum scroll revealed. Nature 626 (7999): 461–62. https://doi.org/10.1038/d41586-024-00346-8.

Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020). How we do things with words: Analyzing text as social and cultural data. Frontiers in Artificial Intelligence, 3, 62. https://doi.org/10.3389/frai.2020.00062.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre*, *Faculty of Informatics*, *Masaryk University*, *Brno*, *Czech Republic*, 3(2).

Underwood, T. (2019). Distant horizons: digital evidence and literary change. University of Chicago Press.

Varga-Dobai, K. (2013). Gender Issues in Multicultural Children's Literature—Black and Third-World Feminist Critiques of Appropriation, Essentialism, and Us/Other Binary Oppositions. Multicultural Perspectives, 15(3), 141-147. https://doi.org/10.1080/15210960.2013.809303.

Walsh, M., & Antoniak, M. (2021). The goodreads “classics”: a computational study of readers, Amazon, and crowdsourced amateur criticism. Journal of Cultural Analytics, 6(2), 243-287. https://doi.org/10.22148/001c.22221.

Wan, M., & McAuley, J. (2018). Item recommendation on monotonic behavior chains. In Proceedings of the 12th ACM conference on recommender, 86–94. Vancouver British Columbia Canada: ACM. https://doi.org/10.1145/3240323.3240369.

Wan, M., Misra, R., Nakashole, N., & McAuley, J. (2019). Fine-grained spoiler detection from large-scale review corpora. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2605–10. https://doi.org/10.18653/v1/P19-1248.

Windhager, F., & Mayr, E. (2024). Digital Humanities and Distributed Cognition: From a Lack of Theory to its Visual Augmentation. Journal of Cultural Analytics, 7(4). https://doi.org/10.22148/001c.121866.

Wu, P. F. (2023). Veni, vidi, vici? On the rise of scrape-and-report scholarship in online reviews research. Journal of the Association for Information Science and Technology, 74(2), 145-149.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Appendix

Appendix 1. The list of words used for retrieving reviews that mention family members

'my mother', 'my mom', 'my father', 'my dad', 'my parent', 'my parents', 'my child', 'my children', 'my kid', 'my kids', 'my baby', 'my babies', 'my son', 'my sons', 'my daughter', 'my daughters', 'my boys', 'my girls', 'my sister', 'my sisters', 'my brother', 'my brothers', 'my twin', 'my twins', 'my sibling', 'my siblings', 'my grandmother', 'my grandmothers', 'my grandma', 'my grandpa', 'my grandfather', 'my grandfathers', 'my grandparent', 'my grandparents', 'my grandson', 'my grandsons', 'my granddaughter', 'my granddaughters', 'my grandchild', 'my grandchildren', 'my aunt', 'my aunts', 'my uncle', 'my uncles', 'my niece', 'my nieces', 'my nephew', 'my nephews', 'my cousin', 'my cousins', 'my husband', 'my wife', 'my partner', 'my spouse', 'my sister-in-law', 'my sisters-in-law', 'my brother-in-law', 'my brothers-in-law', 'my mother-in-law', 'my father-in-law', 'my boyfriend', 'my girlfriend', 'my fiancé', 'my fiancée', 'my younger', 'my older', 'my elder', 'my friend', 'my friends', 'my family'.

Appendix 2. The forty topic models (ten words per topic)

(0, '0.189*"reread" + 0.126*"level" + 0.119*"silly" + 0.117*"perhaps" + 0.069*"indeed" + 0.038*"typical" + 0.034*"intend" + 0.030*"behaviour" + 0.028*"happily" + 0.027*"insist"')

(1, '0.635*"friend" + 0.095*"pass" + 0.084*"positive" + 0.065*"discussion" + 0.048*"appropriate" + 0.000*"teorii" + 0.000*"sursa" + 0.000*"toate" + 0.000*"studiat" + 0.000*"urmaresc"')

(2, '0.238*"movie" + 0.206*"hear" + 0.158*"listen" + 0.131*"adorable" + 0.119*"bedtime" + 0.039*"flow" + 0.022*"listening" + 0.005*"maker" + 0.005*"player" + 0.001*"compilation"')

(3, '0.279*"begin" + 0.146*"gift" + 0.092*"fan" + 0.084*"consider" + 0.083*"enjoyable" + 0.072*"piece" + 0.052*"window" + 0.042*"incredible" + 0.031*"tradition" + 0.020*"outside"')

(4, '0.451*"definitely" + 0.121*"death" + 0.096*"appeal" + 0.084*"though" + 0.070*"nearly" + 0.035*"own" + 0.031*"memorable" + 0.000*"stele" + 0.000*"stiti" + 0.000*"studiat"')

(5, '0.299*"care" + 0.182*"woman" + 0.151*"grandmother" + 0.114*"various" + 0.050*"charm" + 0.039*"separate" + 0.024*"claim" + 0.004*"ostensibly" + 0.000*"sursa" + 0.000*"timpului"')

(6, '0.093*"book" + 0.055*"read" + 0.034*"story" + 0.030*"love" + 0.029*"child" + 0.018*"get" + 0.017*"make" + 0.017*"time" + 0.017*"kid" + 0.015*"go"')

(7, '0.400*"girl" + 0.237*"home" + 0.165*"sister" + 0.078*"smile" + 0.067*"fantastic" + 0.003*"backyard" + 0.000*"teorii" + 0.000*"stiti" + 0.000*"studiat" + 0.000*"sursa"')

(8, '0.202*"night" + 0.191*"watch" + 0.139*"often" + 0.061*"chance" + 0.045*"poem" + 0.041*"modern" + 0.038*"mark" + 0.032*"famous" + 0.032*"wide" + 0.027*"belong"')

(9, '0.301*"school" + 0.068*"sense" + 0.065*"magic" + 0.061*"present" + 0.058*"relate" + 0.041*"middle" + 0.040*"difference" + 0.039*"notice" + 0.028*"culture" + 0.025*"happiness"')

(10, '0.130*"original" + 0.110*"introduce" + 0.090*"edition" + 0.088*"self" + 0.083*"surprise" + 0.077*"hide" + 0.075*"secret" + 0.047*"steal" + 0.045*"fair" + 0.039*"print"')

(11, '0.210*"absolutely" + 0.199*"away" + 0.116*"mystery" + 0.081*"creature" + 0.076*"town" + 0.045*"solve" + 0.032*"bond" + 0.027*"premise" + 0.024*"gentle" + 0.018*"neighborhood"')

(12, '0.098*"sometimes" + 0.078*"least" + 0.066*"free" + 0.060*"guess" + 0.056*"summer" + 0.048*"situation" + 0.047*"power" + 0.039*"shelf" + 0.038*"entertain" + 0.037*"certain"')

(13, '0.082*"illustration" + 0.075*"thing" + 0.067*"look" + 0.060*"page" + 0.056*"picture" + 0.049*"word" + 0.046*"different" + 0.040*"show" + 0.032*"kind" + 0.032*"share"')

(14, '0.461*"mother" + 0.107*"adore" + 0.096*"niece" + 0.057*"forest" + 0.046*"enchant" + 0.040*"fabulous" + 0.031*"originally" + 0.019*"compassion" + 0.018*"naturally" + 0.011*"hood"')

(15, '0.324*"ever" + 0.216*"classic" + 0.129*"delightful" + 0.064*"encourage" + 0.046*"encounter" + 0.043*"sequel" + 0.037*"comment" + 0.037*"look_forward" + 0.017*"lovable" + 0.003*"somehow_misse"')

(16, '0.163*"cat" + 0.156*"teacher" + 0.141*"aloud" + 0.130*"class" + 0.096*"view" + 0.064*"topic" + 0.033*"poetry" + 0.028*"heavy" + 0.027*"th_grade" + 0.023*"center"')

(17, '0.211*"believe" + 0.175*"appreciate" + 0.154*"voice" + 0.130*"plan" + 0.081*"english" + 0.066*"act" + 0.052*"front" + 0.029*"genius" + 0.009*"essentially" + 0.000*"stiti"')

(18, '0.196*"leave" + 0.150*"let" + 0.140*"lose" + 0.136*"bad" + 0.099*"line" + 0.072*"imagine" + 0.045*"scary" + 0.027*"identify" + 0.023*"somewhat" + 0.016*"medium"')

(19, '0.073*"world" + 0.073*"favorite" + 0.061*"live" + 0.052*"grow" + 0.044*"become" + 0.039*"play" + 0.038*"place" + 0.029*"set" + 0.028*"remind" + 0.026*"bear"')

(20, '0.455*"family" + 0.198*"name" + 0.132*"tale" + 0.109*"father" + 0.021*"dance" + 0.017*"orphan" + 0.011*"adopt" + 0.002*"actress" + 0.000*"sursa" + 0.000*"stiti"')

(21, '0.000*"timpului" + 0.000*"toate" + 0.000*"romanian" + 0.000*"sen" + 0.000*"stele" + 0.000*"stiti" + 0.000*"studiat" + 0.000*"sursa" + 0.000*"teorii" + 0.000*"proza"')

(22, '0.116*"big" + 0.108*"fact" + 0.100*"mom" + 0.068*"plot" + 0.053*"soon" + 0.049*"realize" + 0.048*"job" + 0.043*"interested" + 0.025*"car" + 0.024*"awesome"')

(23, '0.121*"true" + 0.088*"imagination" + 0.076*"simply" + 0.062*"main" + 0.056*"grade" + 0.054*"fly" + 0.052*"letter" + 0.052*"top" + 0.052*"actual" + 0.051*"beautifully"')

(24, '0.240*"star" + 0.136*"question" + 0.130*"sweet" + 0.125*"person" + 0.056*"easily" + 0.037*"reviewer" + 0.032*"agree" + 0.029*"study" + 0.027*"college" + 0.024*"humorous"')

(25, '0.292*"wait" + 0.273*"maybe" + 0.131*"hate" + 0.067*"elephant" + 0.061*"egg" + 0.038*"stupid" + 0.016*"crush" + 0.000*"teorii" + 0.000*"toate" + 0.000*"wasteful"')

(26, '0.257*"animal" + 0.105*"detail" + 0.070*"sad" + 0.057*"colorful" + 0.054*"number" + 0.051*"fantasy" + 0.046*"cool" + 0.042*"stick" + 0.035*"count" + 0.030*"sign"')

(27, '0.000*"timpului" + 0.000*"toate" + 0.000*"romanian" + 0.000*"sen" + 0.000*"stele" + 0.000*"stiti" + 0.000*"studiat" + 0.000*"sursa" + 0.000*"teorii" + 0.000*"proza"')

(28, '0.701*"daughter" + 0.079*"relationship" + 0.040*"store" + 0.030*"shoe" + 0.023*"grandpa" + 0.017*"kick" + 0.013*"inappropriate" + 0.008*"dollar" + 0.006*"frustrated" + 0.004*"meanwhile"')

(29, '0.150*"chapter" + 0.122*"move" + 0.111*"face" + 0.075*"light" + 0.059*"fill" + 0.055*"sleep" + 0.043*"city" + 0.037*"country" + 0.036*"door" + 0.036*"ready"')

(30, '0.155*"funny" + 0.148*"lesson" + 0.090*"style" + 0.077*"writing" + 0.072*"illustrate" + 0.065*"student" + 0.047*"description" + 0.038*"similar" + 0.036*"witch" + 0.026*"cousin"')

(31, '0.000*"timpului" + 0.000*"toate" + 0.000*"romanian" + 0.000*"sen" + 0.000*"stele" + 0.000*"stiti" + 0.000*"studiat" + 0.000*"sursa" + 0.000*"teorii" + 0.000*"proza"')

(32, '0.073*"part" + 0.069*"back" + 0.045*"quite" + 0.042*"point" + 0.039*"especially" + 0.039*"together" + 0.036*"perfect" + 0.036*"one" + 0.033*"course" + 0.032*"right"')

(33, '0.160*"message" + 0.160*"sit" + 0.149*"special" + 0.103*"month" + 0.099*"glad" + 0.057*"twist" + 0.052*"nephew" + 0.043*"throw" + 0.038*"approach" + 0.021*"plant"')

(34, '0.231*"wonderful" + 0.098*"baby" + 0.098*"receive" + 0.089*"memory" + 0.072*"favourite" + 0.067*"dream" + 0.048*"hour" + 0.045*"touch" + 0.042*"future" + 0.035*"red"')

(35, '0.431*"boy" + 0.124*"cry" + 0.116*"tree" + 0.080*"tear" + 0.074*"perfectly" + 0.046*"single" + 0.017*"boyfriend" + 0.015*"sharing" + 0.013*"sible" + 0.009*"unconditional"')

(36, '0.290*"short" + 0.159*"bed" + 0.144*"base" + 0.081*"fairy_tale" + 0.075*"lie" + 0.045*"mine" + 0.033*"marry" + 0.028*"flip" + 0.013*"ish" + 0.012*"myth"')

(37, '0.146*"talk" + 0.121*"laugh" + 0.088*"draw" + 0.077*"explain" + 0.069*"several" + 0.061*"loud" + 0.060*"opinion" + 0.047*"alone" + 0.034*"space" + 0.033*"afraid"')

(38, '0.203*"add" + 0.152*"language" + 0.136*"list" + 0.107*"struggle" + 0.101*"strong" + 0.094*"event" + 0.073*"charming" + 0.044*"fiction" + 0.009*"non_fiction" + 0.002*"license"')

(39, '0.142*"art" + 0.096*"black" + 0.085*"build" + 0.082*"princess" + 0.080*"white" + 0.078*"hit" + 0.060*"aspect" + 0.052*"fast" + 0.051*"money" + 0.047*"reach"')

Information Research - Vol. 30 No. iConf (2025)

Conceptual approaches to information-as-potentiality

Alexandra Chassanoff and Annie T. Chen

DOI: https://doi.org/10.47989/ir30iConf47266

Abstract

Introduction. In this paper, we advocate for the development of holistic approaches, theories, and frameworks for recasting the study of information use as a dynamic, emergent process.

Method. We begin by reviewing the LIS literature on information use to demonstrate the limitations of information-theoretical perspectives. We use the exemplar scenario of digital collections to demonstrate the need for open-ended, qualitative approaches for studying information in the context of its use.

Results. We introduce a conceptual approach called '*information-as-potentiality*' and articulate a set of four meta-theoretical lenses (experienced information, situated information, participatory information, and embodied information) that can be used to ground empirical inquiry into the unfolding process of information.

Conclusion. We conclude by reviewing the benefits of adopting '*information-as-potentiality*' perspectives and consider their potential value for informing the design and evaluation of digital collections.

Introduction

Understanding how scholars interpret and use artifacts from digital collections in research pursuits is an essential part of designing and building supportive knowledge infrastructures (Edwards et al., 2013). In these rapidly evolving information environments, interactions are mediated by interfaces, search queries, navigation, classification structures, and other factors that often require specific fluencies and competencies. Observing these new modes of practice in different scholarly domains can generate critical insights about changing epistemic norms and evolving literacies.

Traditionally, libraries and archives/special collections have used methods like circulation analysis and user surveys to determine the value and impact of collection information for users. Evaluating the use of digital collections, however, is difficult because there are currently no standardized assessment measures (Hughes, 2014; Kamposiori, 2020). Empirical studies on digital collections often employ quantified (e.g., transaction logs, web analytics) or causal (e.g., task-based retrieval) approaches, which can fail to capture the complexities of multimodal, multi-stage information processes. Consequently, there is very little qualitative understanding about the organizational, technical, and/or social practices that impede or enhance user interactions in digital collection environments.

In this paper, we propose a holistic approach to studying the epistemic possibilities of information in the context of its use through the conceptual lens of *information-as-potentiality*. We agree with Day (2000, 2011) and more recently Huvila (2018; 2022) that library and information science (LIS) should prioritize the development of methodological frameworks and theories for studying information use as a continuous, highly mediated activity. We use the LIS literature and related work to articulate a set of four meta-theoretical lenses for framing inquiry into information-as-potentiality within the context of evaluating digital collections use. In doing so, we are able to focus attention on the unfolding of dynamic, interpretive information practices and processes that configure meaningful (or meaningless) information use.

Background

Beyond the conduit: information use and digital collections

The information theories developed by Claude Shannon and Warren Weaver in their experiments on communication have had a significant influence on the discipline. Originally published in 1948, Shannon and Weaver's engineering model is primarily concerned with effective message transmission across an information channel and factors influencing message reception (such as channel capacity, noisiness, or degrees of uncertainty). The model is explicit about not incorporating meaning into a definition of information, claiming that the '*semantic aspects of communication are irrelevant to the engineering problem*' (Shannon, 1948).

Scholars from a variety of disciplines have long argued that information theory and its reductions a la the '*conduit metaphor*' have hindered the development of robust understandings of information as a dynamic, emergent process (Reddy, 1977; Soto and Sonnenschein, 2020). LIS (library and information science) does not define concepts, attributes or material properties of information and tends to defer to a larger framework concerned with behavior *around* information. Within the discipline, information-theoretic influences on the operationalization of the field's target value resolve primarily as either painting information use as a causal process of linear transmission and/or portraying information as a variable '*thing*' that can be represented for retrieval and/or preservation. Both conceptions have significant consequences for how we study and understand information in its different contexts of use. Day (2000, p.808) argues that the conduit approach has structured '*how we conceive and define the nature and scope of the field, what methods are allowable, what researchers, researchers, and disciplines are valuable for information studies.*' According to Capurro (1991), proponents of information theory conceive of information as a

singular, fixed representational value that flows linearly between two points. Information thus becomes a one-dimensional message, a '*substance that is transported between the sender and the receiver*' (Mai, 2013, p.676). Consequently, there is very little qualitative understanding about the evidential practices that *make something* information (Furner, 2004).

Cultural heritage institutions have long acknowledged difficulties tracking and evaluating the use of their collections (Dearstyne, 1987; Chapman et al., 2015) and the transformation to digital collections has only amplified an existing problem space for assessment. Namely, what constitutes the use of a [digital] collection? Does clicking on an artifact indicate use? Should it? In LIS studies on digital collections use, the so-called '*conduit approach*' has influenced the kinds of methodologies employed. In the first approach, collection use is quantifiably assessed through measuring access to digital resources. Researchers use transaction logs or web site analytics as tools to evaluate resource use (Prom, 2011; Lapworth, 2020). A second approach assesses use through the occurrence of citations and/or references in publications (Sinn, 2012; Harris and Hepburn, 2013). In the third approach, scholarly preferences for accessing collections are analysed, with an emphasis on whether scholars access materials through print or online methods (Tibbo, 2003; Anderson, 2004; Duff, Craig, and Cherry, 2004).

These studies provide very little insight into the multi-faceted, interactional space of scholarly information use in digital collection spaces. There are inherent limitations with employing quantitatively driven frameworks, theories, and methodologies to frame questions of information use - itself a highly interpretive, and often undefined sphere of information activities. In the first and third approaches, researchers are studying access to (rather than use of) information sources (Spink and Cole, 2006). Citation analysis is only partially helpful as an indicator of use; users might consult multiple sources in the course of their research but may not necessarily cite them. A further problem with citation analysis is the relative newness of standardized practices for citing digital resources, likely painting an inaccurate picture of use. As Sheble and Wildemuth (2013) note, transaction logs '*cannot tell us anything about users' cognitive or affective responses during the system interaction*' (p.174). Treating resource access and retrieval mechanisms simply as a proxy for information use obscures important aspects of the interaction space. How do factors such as digitization quality, the availability of technical metadata, or navigation options impact scholarly interpretation and use? What contextual qualities support meaningful discovery, access, and use of digital collections, and how does the interaction process influence the experience or outcomes of use? Answers to these questions can help inform the successful design of dynamic collection spaces and influence how LIS proceeds with theoretical developments.

Information use as an ongoing interpretive process

The dynamics of information flows within larger computer-mediated communication ecosystems introduce heuristic and practical complexities. Scholars working in digital collections now encounter artifacts through an interface where form, content, and interpretation converge in a particular space-time context (Hedstrom, 2002). This convergence meshes together '*previously separated realms*' (Lowood, 2002, p.9) into singular knowledge representations conditioned by ongoing mediations. As the factors which mediate content change (e.g., device, modality), so will the interpretation. Writing in the 1960s, the sociologist John McHale warned:

> *we are faced…with the emergence of a new and powerful fusion of technological capabilities which not only potentially amplifies our capabilities to deal with our social and physical environment transactions but which by its function as screen, channel and multiplexer of information actually reshapes the information content and perception of society – in ways that our conventional wisdom and traditional institutional means may not be able to foresee, comprehend or effectively control* (p.16-17).

Studies on digital collections use capture a *specific instantiation* of a space-time bound relational link – an enactment of a viewing modality coupled with a representational level of content. But the specificity of this capture can also obfuscate *the process of information*. What gets lost in translation? In studying use in dynamic environments, being able to capture the organic, dynamic, and evolutionary nature of interpretation is essential. Let us consider '*information use*' on social media. What users see on social media such as Twitter (X), Instagram and Tik Tok are algorithmically mediated; for example, Tik Tok's for you page offers content tailored to the user. The videos – Tik Toks – that each person sees, their order, and how they experience (interpret) them differs. In addition, a person's use experience is shaped dialogically in an ongoing interpretive process, as future content is based on users' past and current interactions with content. Tik Tok users are aware of the impacts of the algorithmic nature of the platform, such as the creation of filter bubbles and the promotion of potentially harmful content, and users change their behaviours to reflect how they understand themselves (Karizat et al., 2021). User behavior, then, can only ever be understood in the context of how they perceive system behavior, themselves, and indeed, even their views of the human experience.

As a user, different social technologies may be associated with different meanings, e.g., Facebook as a space for presentation of the socially acceptable; Instagram, for stylized self-presentation; and Twitter as a venue for informal information exchange (Boczkowski et al., 2018). Companies and users co-create experiences with these platforms over time through features with socially embedded meanings, such as retweets on Twitter (X) and stories on Instagram. Considering users' interactions in context, and through the language and culture of the media can facilitate a more nuanced understanding of meaning-making and use within different social media platforms.

Conceptualizing information-as-potentiality

In this paper, we argue that LIS should develop and adopt alternative frameworks, methodologies, and theories to facilitate the qualitative exploration of information use. We now explore how LIS research can incorporate *information-as-potentiality* in both design and evaluation efforts, with emphasis on digital collections as exemplar scenarios where acts of interpretation continuously mediate use.

Recent LIS studies highlight the importance of grounding study in the activities that constitute and bring forth informational aspects of interactions, what Huvila (2021) calls information making and taking. To this end, we might incorporate user-centred design methods to study interactions with information and technologies. Website visitors' actions might be studied in behavioral analysis, through page views, queries, scrolling, viewing, and view duration (Barifah et al., 2020). Eye tracking might be used to understand user experiences in digital libraries (Kuhar & Mercun, 2022). The think-aloud protocol is often used in interviews to study users' experiences and satisfaction with digital collections (Du et al., 2021; Zahidi et al., 2014). Standardized measures to assess usability such as the User Experience Questionnaire (Laugwitz et al., 2008) and AttrakDiff (Hassenzahl et al., 2003), experience measures the attractiveness of products (Walsh et al., 2014), complement these methods by providing information about user perceptions. These methods are congruent with the capture of motivation, affect, and satisfaction in the study of the use of digital collections (Kuhar & Mercun, 2022), but also afford a richer exploration of information experiences,

Kreiseler et al. (2017) argue that four functional principles support exploration in digital collections – view, movement, contextualization, and participation – and that the exploration of digital collections and cultural artifacts more broadly, could be studied through a three-part methodology, '*reverse information architecture*', that facilitates connection between visual interfaces and design concepts they are based on. Concretely, this involves analysis of navigation, wireframe (page layout) and position (positioning of text and image elements). The idea that information experiences can be modelled applies not only to visual interfaces, but also those involving other modalities.

Practice-based perspectives provide a useful lens for understanding collectively how and why people do things with information (Lloyd, 2010) and can be useful for explicating shared meanings and social contexts underlying sites of interrogation (Schatski, 1995). Concepts relating to visual information practices and knowledge organization, mixed with user-centred methodologies, can also inform design and evaluation efforts. For example, Morse et al. (2021) incorporate the concept of *rich prospect*, an interface design framework that includes seven core principles: representation, organization, depth, availability, multiplicity, coherence, and selection, which was developed to enhance browsing and discovery in visual information collections. They employ rich-prospect in a mixed-methods study of three different cultural heritage collections, *Coins*, *Curator Table*, and *Museum of the World*. After an initial exploratory session using the think-aloud protocol, participants were introduced to the principles of rich-prospect on cards and then asked to consider the collections given these principles and their utility for their browsing experience. The use of these principles, along with other data collection techniques, facilitated a nuanced understanding of the relative contributions of different facets to one's overall experience, which in turn could inform prioritization and shaping of collection features.

Using a synthesis of the literature and our related work, we now introduce a set of meta-theoretical lenses that can guide analytical approaches for designing and evaluating digital collections use. We argue that each of these lenses provide a useful methodological orientation, framing inquiry into facets of informational interactions grounded in critical contexts of use. Metatheories are '*a set of assumptions, epistemological and ontological, that orient and direct theorizing about a given phenomenon*' (Lawler and Ford, 1993, p.4). They can serve as a clarifying device for a disciplinary field to showcase underlying philosophical or theoretical assumptions commonly in play. According to Bates, metatheories represent '*the philosophy behind the theory, the fundamental set of ideas about how phenomena of interest in a particular field should be thought about and researched*' (2005, p. 257). Each lens might be applied separately or together as they are not mutually exclusive, and one or more of them may apply in any given situation. Below, we briefly describe each lens and provide examples from the literature demonstrating their potential applicability in framing inquiry into digital collections use.

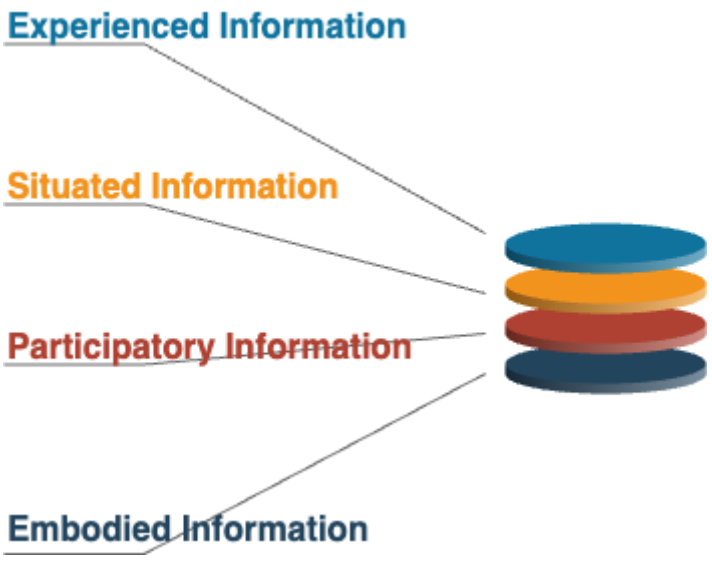


Figure 1. The metatheoretical lenses of *information-as-potentiality*

Meta-theoretical lens #1: experienced information

The lens of experienced information provides a conceptual window into environments where people build understanding by engaging with documentation (Gorichanaz, 2017). Focusing on experienced information emphasizes the interactions between individuals and their life-worlds rather than only capturing the outcomes of those interactions (Bruce, 2014). Holistic accounts that capture how information is experienced prompt attention to processes underlying '*what information a given artifact imparts*' (Hoyte, 2019, p.415).

Information pathways is one methodological approach for studying how information is experienced and has been proposed as a guiding metaphor from which to characterize the design of cultural heritage collections. Goodale et al. (2012) observe that information pathways can have various conceptualizations, including as information-seeking journeys, a starting point or way in, a route through, a learning process, and others. Wray et al. (2013) emphasizes two dimensions of pathways, one involving creative, non-linear journeys (e.g., a search history) and one supporting learning and meaning making (e.g., linked metadata). Assessment of the similarity between works in a collection affords a sense of the potential for non-linear serendipitous explorations. In a similar vein, Adams, and Blandford (2005) performed a grounded theory analysis of the *information journeys* taken by users of digital libraries in two domains, health, and academia, identifying three main stages – information initiation, facilitation (or gathering) and interpretation – and opportunities to better support users in these experiences.

Though log data collected concerning digital collection use are often quantitative in nature, they can be considered more interpretively, as pathways for understanding how information is experienced. Behavioral metrics could serve as a temporal view of the path that one takes – their experience – of an artifact or collection. Though the manifest meaning of page views, click-throughs, and other user interactions may be unclear, an alternate way of conceptualizing log data are as the artifacts of an experiential process. When one views the traces of activity through a website, one is viewing a process of sense-making. Considering users' behaviours sequentially could lead to informative re-constructions of how users learn and engage in meaning-making with collections, as well as paths that perhaps designers had not anticipated, facilitating the imagining of new potentialities.

Meta-theoretical lens #2: situated information

Given the right conditions and environments, all kinds of entities (artifacts, people, events) can inform or '*make something known*' (Briet, 1951), but it is the processing of that entity for '*informational purposes*' (Buckland, 2001) that signifies a transition to documentation status. Taylor (1986;1991) is one LIS scholar who recognized the importance of approaching information from a situated perspective, arguing that specific meaning(s) and value are established in the context of use. He developed a conceptual framework for studies on information use environments, aiming to capture influential elements that impact flow and determination of informational value. According to Taylor, information always exists as a phenomenon of study in a situational context. He explains, '*Data becomes information when we establish relationships among data...it is the structure and relations among data, and most important, the rules and conditions used to establish those relations*' (1991, p. 7-8). Viewed through this lens, artifacts can serve as prompts for informational study and in practice approaches, as '*epistemic objects*' always in the process of unfolding (Knorr-Cetina, 2000). The situated nature of information interactions was also taken up in LIS by Blandford and Attfield (2010), elaborating on previous work theorizing information journeys. They describe how sites of situated information interactions are mediated contexts that are '*physical, social, and activity-centred*' and can overlap situationally (2010, p.2).

In affiliated knowledge communities, the intersecting spheres of information studies, human-computer interaction, and organizational work studies have produced relevant theoretical work on '*the situated entanglement of the technical and the social*' (Orlikowski, 2005, p.185). Interrogating

processes of information from this perspective can be fruitful for unpacking the 'multiple logics that configure together social and material realities (Barad, 2007). Technology itself can be interpreted as a social practice and efforts to study knowledge-generative environments must account for the '*constitutive, co-dependent relations of complexity*' that underpin all digital information environments (Drucker, para. 30, 2013). Studying conditions of situated information permits attention to the '*the individual work and its activation of the particular conventions as its point of departure*' (Doane, p. 131). Rather than focusing on the medium and its corresponding material presence, though, information is studied in the context of its performative becoming; though not the sole, liminal focus of interest, the medium is an inherent part of the (informational) context.

While in principle, studying situated information or information in its '*performative becoming*' may seem intuitive, questions may arise concerning how this might be achieved in practice. There are methods from different disciplines and approaches that can be helpful. For example, the concept of *information horizon* introduced by Sonnenwald (1999), referring to the perceived horizon of potential information resources, affords an intuitive representation of the socially constructed and situated use of information by different users. Often used in contextual inquiry (Beyer & Holtzblatt, 1999), visual tools such as information flow diagrams, cultural models, sequence models and other diagrams might also be used to understand and envision potential information uses. Jo Bates et al. (2016) introduces a methodological approach they call *data journeys* for exploring the socio-material, constitutive nature of digital data in the context of its use. The value of this situated perspective is the ability to attend to *how* information flows through complex environments, making visible '*potential movement, blocked movement and lack of movement.*'

Meta-theoretical lens #3: participatory information

In this paper, we argue that participatory approaches to information practice research can be an effective means of contending with '*information-as-potentiality*' and meaning-making. Let us consider the scenario of information creation and use in heritage sites such as archives and museums. Rather than viewing these as static, participatory information practices invite people to interact with sites as a 'means for social interaction, participation, co-creation and contribution' (Stuedahl, 2018, p. 219). Incorporating users' information practices and potentially leveraging available technologies to assess feelings and reactions in the moment (e.g., through ecological momentary assessment) could facilitate richer engagement with and design of heritage environments. For example, the Acropolis Museum invited museum visitors to evaluate mobile-based storytelling guides designed to facilitate personalized experiences (Roussou & Katifori, 2018). They incorporated data collected through ethnography (i.e., shadow observation), interviews, questionnaires, and data logging into analysis, and ultimately considering themes such as interactive story plot and narration, staging, and wayfinding, personalisation, and social interaction.

Participatory information approaches such as decentralized curation focus on usability and findability ('*radical user orientation*') and contextualisation of both records and the entire archival process (Huvila, 2008); and engaging community members in the appraisal, arrangement, and description phases of creating an archival record, are needed to ensure meaningful representation (Shilton & Srinivasan, 2007). Examples of participatory archives projects are #NorthVanStories, a social media collection project in which residents of North Vancouver were encouraged to tag social media with this hashtag, and #NorthVanStories–Living Histories, a digital video oral history project (Bushey, 2023). #NorthVanStories–Living Histories involved multiple steps, including identification of community assets, selection of projects relevant to the communities, contact with storytellers, sharing a questionnaire with questions to support the aims of the project, and creation of five videos with storytellers from diverse and vulnerable communities. This multi-step process illustrates consideration of the information to be included and scaffolding (the questionnaire) to

support, but not define the knowledge representation, thus preserving voice and flexibility for storytellers to tell stories their own way. In both projects, meaning was defined and evolved through dialectical engagement, thus highlighting the need to: a) design with the participation of communities involved; and b) develop designs that allow for information to be defined and evolve as it is being produced. Thus, participatory information practices facilitate the realization of collective negotiation of the meaning of information as an emergent process.

Meta-theoretical lens #4: embodied information

Information can become embodied in and through a variety of forms that extend beyond the human body. Different contextual factors catalyse the informational process, including affective motivations and goals (Nahl, 2007) and align with domain-specific meanings (Hjorland & Alberton, 1995). Working in the vein of digital information studies, Hayles (2002) argues that the materiality of information arises from the combination of user interactions and interpretive strategies. In other words, information emerges in its material form through the coupling of interpretation with interaction. Dourish (2001) also proposed a theory of embodied interaction that emphasizes context-aware approaches to studies of social interactions with physical and/or technical artifacts. Bates (2018) helpfully reminds us that as humans '*operating in a world of possible affordances*' (p.255) we have the embodied capacity to perceive, experience, and communicate information in meaningful ways.

The relational links between perception and embodiment were made in early critiques of information theory, initially by the British scientist Donald MacKay who noted the significant role of perception in the transfer of information. For MacKay, the act of perceiving messages is made possible by an existing internal state of conditional readiness. He writes,

> *the mechanical energy of a message must be sufficient to do the mechanical job that eventually resets the brain-levels; but the selective job, of determining which levels shall move, depends on the form of the message, and on the state of your brain before you hear it* (1969, p.24).

Drawing attention to the job of selection is critical for MacKay's argument; while all sorts of stimuli might be received, only stimuli that evoke a matching response in the organism can be perceived. This distinction is critical to understanding how perception itself is embodied in the activity of information, or, as MacKay puts it, the '*activity of knowing is an activity of internal adaptive response*' (p. 61).

More recently, the concept of embodied information has received attention by LIS, often by way of studies on information practices. According to Olsson and Lloyd (2016),

> *embodied information practices (social, physical, and material) therefore act as complex and rich sources of information that are referenced through the social, physical, and material conditions of the site and its ways of knowing.*

We also observe embodied ways of knowing in serious leisure, including in ultra-running (Gorichanaz, 2015) and enthusiast car restoration (Lloyd-Zantiotis & Olsson, 2017). Similarities in these examples include the complexities of information practices, their social and oral nature, and the important role of acquired knowledge. However, there are also differences in individual ways of knowing. Whereas, in the context of ultra-running, the body was a key source of information (Gorichanaz, 2015), a participant in Lloyd-Zantiotis and Olsson's (2017) study developed a feel for different types of materials and how to work with them in the context of the car restoration process.

Studies of embodied information have employed a variety of methods, such as interviews, diaries, and observation, that might be useful for capturing the richness of information behavior that occur

– '*information-as–potentiality*' – with digital collections. For example, the experience of historical artifacts in digital form, can be embodied. Green (2018) writes that social media affords '*curatorial hands*' through haptic experiences of digitized manuscripts. Though even the physical access to artifacts is often limited to a select few who are able to travel to view artifacts behind glass display cases, today social media such as Instagram can afford novel forms of engagement manipulating a digital representation via tapping, clicking or swiping, or scrolling back and forth through an set of images curated as a story, thus enabling viewers to interact with artifacts in ways not hitherto possible using traditional viewers. How is such engagement to be measured? One might conceive of the utilization of different ethnographic methods, but also the development of novel instruments that incorporate consideration of the embodied nature of information.

Media archaeology is one possible theoretical approach for augmenting embodied information research in interacting with digital collections. For example, a media archaeological approach could prove useful in presenting scholars with technical (e.g., EXIF) metadata alongside the display of digital artifacts from online collections. How do scholarly interpretations of digital artifacts change in these different contexts? How might such factors inform or contribute to the design and use of digital collections?

Conclusion

This paper is an initial attempt to offer a set of metatheoretical lenses for conceptualizing *information-as-potentiality*. We use the exemplar scenario of digital collections to show how the application of information-theoretical understandings offers an incomplete vantage point for assessing informational qualities generated by interactions, and grounding inquiry in perspectives that frame *information-as-potentiality* cultivates the appropriate, relevant logic from which to evaluate the design and use of digital collections. Conceptualizing information use as an experienced, participatory, situated, and/or embodied set of practices captures key insights into the dynamic, unfolding nature of the information process in the context of use. We offer these conceptual lenses as ways that can help direct our attention to dimensions of interest throughout the information use process, an increasingly important endeavour considering how often we now encounter information across a variety of spatiotemporal dimensions and contexts.

In this paper, we have shown that there is both opportunity and a need to consider the potentiality of information in research involving digital collections and cultural heritage. The use and interpretation of information can be complex, non-linear, situated, and organic, dependent on and evolving alongside user interactions in different contexts. Ultimately, we aim to contribute to a nuanced understanding of how information behaves in computationally aware platforms, interfaces, and networked environments. Though user-centred design, ethnography, and other approaches provide a rich set of methods to study human experience, we suggest that information use studies engaging with the meta-theoretical lenses specified here can further enrich our understanding of dynamic information use behavior and in turn, inform the design of more fulfilling digital collections and other information environments.

Acknowledgements

Alexandra Chassanoff University of North Carolina Chapel Hill and Annie T. Chen University of Washington

About the authors

Alexandra Chassanoff is Assistant Professor in the School of Information and Library Science, University of North Carolina, Chapel Hill, North Carolina, USA. Her research focuses on born-

digital curation and community-driven approaches to cultural heritage. She can be contacted at: achass@unc.edu

Annie Chen is Associate Professor in the Department of Biomedical Informatics and Medical Education at the University of Washington School of Medicine. Her research interests include information behaviours in the context of long-term health management; information and interaction in online communities and other digital spaces; the use of natural language processing and visualization methods to study linguistic patterns relating to psychosocial and communicative processes; and user-centred design. She can be contacted at atchen@uw.edu

References

Adams, A. & Blandford, A. (2005). Digital libraries' support for the user's 'information journey.' *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, 160–169. https://doi.org/10.1145/1065385.1065424

Anderson, I. (2004). Are you being served? Historians and the search for primary sources. *Archivaria* 58, 88–129.

Barifah, M., Landoni, M., & Eddakrouri, A. (2020). Evaluating the user experience in a digital library. *Proceedings of the Association for Information Science and Technology*, 57(1), e280. https://doi.org/10.1002/pra2.280

Bates, M. J. (2005). An introduction to metatheories, theories, and models. In K. E. Fisher, S. Erdelez, & L. McKechnie (Eds.), *Theories of information behavior* (pp. 1–24). Medford, NJ: American Society for Information Science and Technology.

Bates, M.J. (2018). Concepts for the study of information embodiment. *Library Trends* 66(3), 239-266.

Bates, J., Lin, Y.W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data and Society* 3(2).

Beyer, H. & Holtzblatt, K. (1999). Contextual design. *Interactions*, 6(1), 32–42.

Blandford, A. & Attfield, S. J. (2010). *Interacting with information.* Morgan & Claypool Publishers.

Boczkowski, P. J., Matassi, M., & Mitchelstein, E. (2018). How young users deal with multiple platforms: The role of meaning-making in social media repertoires. *Journal of Computer-Mediated Communication*, 23(5), 245–259. https://doi.org/10.1093/jcmc/zmy012

Bruce, C., Davis, K., Hughes, H., Partridge, H. & Stoodley, I. (2014). *Information experience: Approaches to theory and practice* (*Library and Information Science*, Vol. 9), Emerald Group Publishing Limited, Leeds, 3-15. https://doi.org/10.1108/S1876-056220140000010001

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science*, 42(5), 351-360. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

Bushey, J. (2023). A participatory archives approach to fostering connectivity, increasing empathy, and building resilience during the COVID-19 pandemic. *Heritage*, 6(3), Article 3. https://doi.org/10.3390/heritage6030125

Capurro, R. (1991). Some critical comments on three leading paradigms of information science section, para.1," *Foundations of Information Science: Review and Perspectives*. https://www.capurro.de/tampere91.htm

Chapman, J., DeRidder, J. & Thompson, S. (2015). Developing best practices in digital library assessment: Year one update. *D-Lib Magazine*, 21, 11-12. https://www.dlib.org/dlib/november15/chapman/11chapman.html

Day, R.E. (2000). The "conduit metaphor" and the nature and politics of information studies. *Journal of the American Society for Information Science & Technology*, 51(9), 805-811.

Day, R.E. (2011). Death of the user: Reconceptualizing subjects, objects, and their relations. *Journal of the American Society for Information Science & Technology*, 62(1), 78-88.

Dearstyne, B. (1987). What is the *use* of archives? A challenge for the profession. *The American Archivist*, 50(1), 76-87. https://doi.org/10.17723/aarc.50.1.572q383767657258

Dourish, P. (2001). Seeking a foundation for context-aware computing. *Human–Computer Interaction*, 16(2), 229–241. https://doi.org/10.1207/S15327051HCI16234_07

Drucker, J. (2013). Performative materiality and theoretical approaches to interface. *Digital Humanities Quarterly*, 7(1). https://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html

Du, J., Yuen, C., Slaughter, M., & Chen, A. T. (2021). Perceived usability and experience with digital tools in the context of digital humanities research. *Proceedings of the Association for Information Science and Technology*, 58(1), 435-439. https://doi.org/10.1002/pra2.474

Duff, W., Craig, B., & Cherry, J. (2004). Finding and using archival resources: A cross-Canada survey of historians studying Canadian history. *Archivaria*, 58, 51–80.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013). Knowledge infrastructures: Intellectual frameworks and research challenges. University of Michigan. http://hdl.handle.net/2027.42/97552

Furner, J. (2004). Information studies without information. *Library Trends*, 52(3), 427-446.

Goodale, P., Clough, P., Ford, N., Hall, M., Stevenson, M., Fernando, S., Aletras, N., Fernie, K., Archer, P., & de Polo, A. (2012). *User-centred design to support exploration and path creation in cultural heritage collections*, 75–78. https://oro.open.ac.uk/68751/

Gorichanaz, T. (2015). Information on the run: Experiencing information during an ultramarathon. *Information Research: An International Electronic Journal*, 20(4). https://eric.ed.gov/?id=EJ1087433

Green, J. M. E. (2018). Digital manuscripts as sites of touch: social media for "hands-on" engagement with medieval manuscript materiality. *Archive Journal*. www.archivejournal.net

Harris, V. & Hepburn, P. (2013). Trends in image use by historians and the implications for librarians and archivists. *College and Research Libraries*, 74(3).

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus & J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-80058-9_19

Hedstrom, M. (2002). Archives, memory, and interfaces with the past. *Archival Science*, 2(1-2), 21–43. https://doi.org/10.1007/BF02435629

Hjørland, B. and Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, 46(6), 400-425.

Hoyte, P. (2019), Implications of the embodied, enactive mind on theorizing about information experience. *Proceedings of the Association for Information Science and Technology* 56, 413-416. https://doi.org/10.1002/pra2.40

Huvila, I. (2008). Participatory archive: Towards decentralised curation, radical user orientation, and broader contextualisation of records management. *Archival Science*, 8(1), 15–36. https://doi.org/10.1007/s10502-008-9071-0

Huvila, I. (2018). Ecology of archaeological information work. In I. Huvila (Ed.), *Archaeology and archaeological information in the digital society* (pp. 121–141). Routledge.

Huvila, I. (2022). Making and taking information. *Journal of the Association for Information Science and Technology*, 73(4), 528–541. https://doi.org/10.1002/asi.24599

Kamposiori, C. (2020). Measuring the impact of special collections and archives in the digital age: Opportunities and challenges. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 30(1), 1-31. https://doi.org/10.18352/lq.10345

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1–44. https://doi.org/10.1145/3476046

Knorr-Cetina, K. (2000). Objectual practice. In Karin Knorr Cetina, Theodore R. Schatzki & Eike v. on Savigny (Eds.), *The practice turn in contemporary theory.* New York: Routledge.

Kreiseler, S., Brüggemann, V., & Dörk, M. (2017). Tracing exploratory modes in digital collections of museum web sites using reverse information architecture. *First Monday*. https://doi.org/10.5210/fm.v22i4.6984

Kuhar, M. & Merčun, T. (2022). Exploring user experience in digital libraries through questionnaire and eye-tracking data. *Library & Information Science Research*, 44(3). https://doi.org/10.1016/j.lisr.2022.101175

Lapworth, E. (2021). Assessing large-scale digitization using web analytics. *Digital Library Perspectives*, 37(2), 133-150. https://doi.org/10.1108/DLP-09-2020-0095

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *HCI and Usability for Education and Work* (pp. 63–76). Springer. https://doi.org/10.1007/978-3-540-89350-9_6

Lawler, E. J., & Ford, R. (1993). Metatheory and friendly competition in theory growth: The case of power processes in bargaining. http://digitalcommons.ilr.cornell.edu/articles/1195

Lloyd-Zantiotis, A., & Olsson, M. (2017). Being in place: Embodied information practices. *Information Research: An International Electronic Journal*, 22(1), Article 1. http://informationr.net/ir/22-1/colis/colis1601.html

Lowood, H. (2002). Shall we play a game: Thoughts on the computer game archive of the future. *Bits of Culture*. Stanford University. https://web.stanford.edu/~lowood/Texts/shall_game.pdf

MacKay, D. (1969). *Information, mechanism and meaning*. Cambridge: MIT Press.

McHale, J. (1976). *The changing information environment.* Boulder: Westview Press.

Mai, J.E. (2013). The quality and quantities of information. *Journal of the American Society for Information Science and Technology*, 64(4).

Morse, C., Niess, J., Lallemand, C., Wieneke, L., & Koenig, V. (2021). Casual leisure in rich-prospect: Advancing visual information behavior for digital museum collections. *Journal on Computing and Cultural Heritage*, 14(3), 1–23. https://doi.org/10.1145/3437257

Nahl, D. (2007). The centrality of the affective in information behavior. In D. Nahl & D. Bilal (Eds.), *Information and emotion: The emergent affective paradigm in information behavior research and theory*, Medford, New Jersey: American Society for Information Science and Technology

Olsson, M. & Lloyd, A. (2017). Being in place: embodied information practices. *Information Research*, 22(1), CoLIS paper 1601. http://InformationR.net/ir/22-1/colis/colis1601.html

Orlikowski, W.J. (2005). Material works: Exploring the situated entanglement of technological performativity and human agency. *Scandanavian Journal of Information Systems* 17(1), Article 6.

Prom, C.J. (2011). Using web analytics to improve online access to archival resources. *American Archivist* 74(1), 158–84.

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 284–310). Cambridge: Cambridge University Press.

Roussou, M. & Katifori, A. (2018). Flow, staging, wayfinding, personalization: Evaluating user experience with mobile museum narratives. *Multimodal Technologies and Interaction*, 2(2), 32. https://doi.org/10.3390/mti2020032

Shannon, C.E. (1948). The mathematical theory of communication. *The Bell System Technical Journal* 27, 379-423.

Sheble, L. & Wildemuth, B. (2009). Transaction logs. In B. Wildemuth (Ed.), *Applications of social research methods to questions in Information and Library Science* (pp. 166-177). Westport, CT: Libraries Unlimited).

Shilton, K., & Srinivasan, R. (2007). Participatory appraisal and arrangement for multicultural archival collections. *Archivaria*, 87–101.

Sinn, D. (2012). Impact of digital archival collections on historical research. *Journal of the American Society for Information Science and Technology*, 63(8), 1521-1537.

Sonnenwald, D. (1999). *Evolving perspectives of human information behavior: Contexts, situations, social networks, and information horizons*. In Wilson, T. & Allen, D. (Eds.). *Exploring the contexts of information behavior* (pp. 176-190). London: Taylor Graham.

Soto, A.M. & Sonnenschein C. (2020). Information, programme, signal: Dead metaphors that negate the agency of organisms. *Interdisciplinary Science Review*, 45(3), 331-343.

Spink, A. & Cole, C. (2006). Human information behavior: Integrating diverse approaches and information use. *Journal of the American Society for Information Science and Technology* 57(1), 25-35.

Stuedahl, D. (2018). Participation in design and changing practices of museum development. In *The Routledge Handbook of Museums, Media, and Communication* (pp. 219-231). London: Routledge Publishing.

Taylor, R. (1986). *Value-added processes in information systems*. New Jersey: Ablex Publishing Corporation.

Taylor, R. (1991). Information use environments, In B. Dervin & M.J. Voight (Eds.), *Progress in Communication Sciences* Vol. 10. Norwood, NJ: Ablex Publishing Corp.

Tibbo, H. (2003). Primarily history in America: How US historians search for primary materials at the dawn of the digital age. *The American Archivist*, 66(1), 9–50.

Walsh, T., Varsaluoma, J., Kujala, S., Nurkka, P., Petrie, H. & Power, C. (2014). Axe UX: Exploring long-term user experience with iScale and AttrakDiff. *Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services*, 32–39. https://doi.org/10.1145/2676467.2676480

Wray, T., Eklund, P. & Kautz, K. (2013). Pathways through information landscapes: alternative design criteria for digital art collections. *International Conference on Information Systems* (pp. 1-22). United States: Association for Information Systems.

Zahidi, Z., Lim, Y. P., & Woods, P. C. (2014). Understanding the user experience (UX) factors that influence user satisfaction in digital culture heritage online collections for non-expert users. 2014 *Science and Information Conference*, 57–63. https://doi.org/10.1109/SAI.2014.6918172

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

An analysis of poet demographic and thematic diversity in a poetry collection for inclusive AI

Kahyun Choi and Gyuri Kang

DOI: https://doi.org/10.47989/ir30iConf47263

Abstract

Introduction. AI technologies, such as theme classification and named entity recognition, enhance digital library accessibility. However, they may introduce biases if training datasets lack adequate representation. For instance, prior AI models for poetry classification overlooked dataset diversity, raising concerns about representation. To address this issue, this study assesses the dataset representation and examines potential issues in AI model design for poetry collections.

Method. We annotated and published the race and ethnicity of poets in an American poetry collection curated by *poets.org*, which was recently used to train a poetry theme classification system. We then examined the diversity of the collection using these annotations.

Analysis. We compared the racial/ethnic composition of the collection to U.S. Census data and conducted group-exclusive top word analysis, popular theme analysis, and entropy-based analysis of theme distribution diversity to evaluate linguistic and thematic diversity.

Results. Our findings indicate that most underrepresented groups are well-represented in the collection, except for Latino/a/x American poets. Furthermore, we found that poems from underrepresented groups increase the collection's linguistic and thematic diversity.

Conclusions. To design responsible AI that embraces diversity, it is essential to assess dataset representation and support non-standard English and diverse themes beyond those popular with the general population.

Introduction

Artificial intelligence (AI) has unlocked the potential for enhanced recommendation and search services within online library collections. However, careless use of AI can introduce pitfalls, such as biases embedded in the collections upon which AI models are based, leading to the models' biased outcomes (Cordell, 2020). Indeed, collections often do not represent the general population accurately, causing models to discriminate against marginalized groups (D'ignazio & Klein, 2023). For instance, the recent large language model, GPT-3, demonstrated biases against Muslims and other marginalized groups (Bommasani et al., 2021). Thus, investing extra effort and attention in data collection for AI models, with an emphasis on equity and representation, is crucial (Jo & Gebru, 2020). Moreover, it is critical to assess the equity and diversity of pre-existing collections before use.

Addressing this challenge becomes urgent in the poetry domain as readership rises and the application of AI in poetry analysis expands. National endowment for the arts survey has revealed a 76% increase in US poetry readership from 2012 to 2017 (Iyengar et al., 2018). A follow-up survey in 2022 shows a similarly high engagement, as 11.5% of US adults engage with poetry through reading or listening (Iyengar, 2023). Alongside, various AI systems for improving accessibility of poetry collections, such as poetry theme classification have been developed (Rakshit et al., 2015; Lou et al., 2015; Kaur & Saini, 2017; Navarro-Colorado, 2018; Choi, 2023). Yet, these studies have not examined if the collections were diverse enough to accurately represent the general population. Thus, it remains uncertain whether they adequately account for poems by poets from underrepresented groups.

As part of the '*unbiased AI for poetry analysis: toward equitable and diverse digital libraries*' project funded by institute of museum and library services (IMLS), our study addresses this oversight by assessing *poets.org*'s curated poem collection, with a focus on the race and ethnicity of the poets. We selected this collection because it was used to train a theme classification system in one of the most recent AI systems for poetry (Choi, 2023). Specifically, we compared the U.S. Census population data on race and ethnicity with those of our poem collection to determine if the collection accurately reflects the general population. Furthermore, we analysed prevalent words and themes in poems written by these groups. We investigated how underrepresented groups' work contribute to the word and theme diversity of the poem collection. Also, we identified potential issues with AI systems that are developed mostly based on poems by dominant groups, while works from underrepresented groups are disregarded as outliers. While our primary focus is on this poetry collection, we suggest that our methodology for assessing demographic representativeness of a collection can be applied to other literary genres.

Diversity analysis

Data collection and pre-processing

We utilized the same collection of poems retrieved in October 2022 that Choi (2023) used to train their theme classification system. The poems are from *poets.org*, which is managed by the Academy of American Poets, the nonprofit charitable organization dedicated to fostering American poets and poetry. This collection features a wide variety of poems, each accompanied by several descriptive tags. We excluded audio-only entries, duplicates, and excessively short, resulting 9,445 works. As our project focuses on modern and contemporary American poetry, we selected 8,912 poems published after 1890, covering the era from Emily Dickson and Walt Whitman to 2022. The list of poets, categorized by race and ethnicity, can be available via this link: https://doi.org/10.6084/m9.figshare.25572459.v1

We have identified the racial and ethnic groups of poets using the '*occasions*' field in the `poets.org` collection, especially when the occasion is relevant to specific racial and ethnic groups.

To align with US Census racial and ethnic categories, we selected the following occasions: 'Asian/Pacific American Heritage Month' for Asian/Pacific Americans (APA), 'Black History Month' or African Americans (AA), 'Hispanic Heritage Month' for Latino/a/x American (LXA), and 'Native American Heritage Month' for Native Americans (NA). To better follow US Census categories, we distinguished between Asians and Pacific Islanders by manually identifying poems by Pacific Islanders through a review of the descriptions of the poems. Therefore, in this paper, APA-AA represents Asian Americans without Pacific Islanders, while APA-PA denotes Pacific Islanders exclusively. In our study, 'Others' or 'General' poems represent works by poets who are not associated with the specific occasions we use to identify the underrepresented groups. Additionally, we have reassigned 329 poets to specific underrepresented groups after reviewing their biography pages on `poets.org`; however, these biographies are not detailed enough to distinguish the mixed-race identities recognized by the US Census. Upon determining the racial and ethnic categories of the poets, we organized their poems into the corresponding categories. However, we acknowledge that our grouping strategy may overlook some poems by underrepresented groups, despite its comprehensiveness.

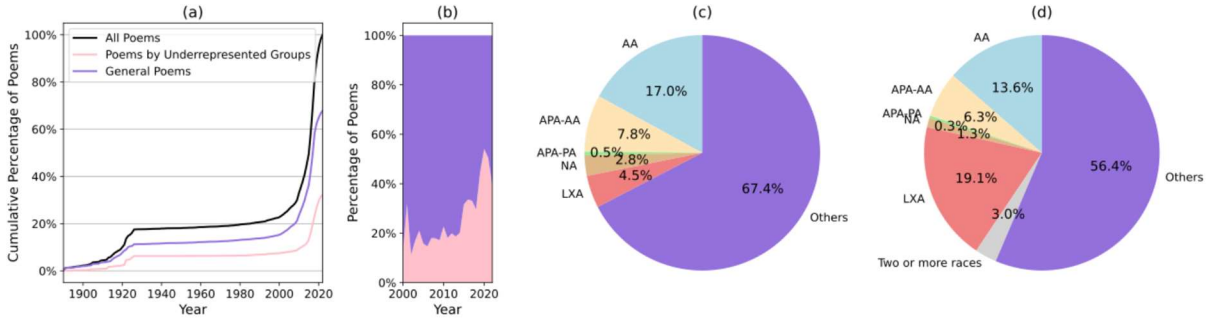


Figure 1. Representation and trends of poet groups in the United States: a) cumulative count of poems; b) yearly ratio of poems by underrepresented groups and general poets since 2000; c) composition of poems by group; d) U.S. population composition

Representation analysis of poems by groups

We examined the cumulative percentage of poems written by poets from underrepresented groups, by other poets, and both combined over time (see Figure1-a). The year indicates either the date of original publication or the date when *poets.org* made the work available on their website. In the collection, there is an increase until mid 1920s, followed by a plateau until early 2000s, and then a soaring upward trend. This pattern can be understood in the context of copyright laws; poems published before 1929 are copyright-free and can be published without permission, while those released after 1929 can only be published on the web with permission. Since its launch in 1996, `poets.org` has published approximately 80% of poems that are copyright-protected with permission. The scarcity of poems between the mid-1920s and early 2000s is attributed to the copyright issues or potential overriding of original publication year by the year of publication on `poets.org`. Over the last two decades, we can observe rapid growth in both general poems and those by underrepresented groups; notably, the latter have exhibited a sharp upward growth trend in the last decade, with recent years showing their proportion to be about half (see Figure 1-b), indicating *poets.org*'s commitment to equity and diversity.

Furthermore, we compare the composition of poems by individual groups with U.S. population demographics (U.S. Census Bureau, 2022) to evaluate how well poems from underrepresented groups are represented in the dataset. Figures 1-c and 1-d show that most underrepresented groups have higher representation, except for LXA: while the LXA population in the U.S. accounts for 19.1%, their representation in the collection is only 4.5%. This gap might be due to the prevalence of Spanish in the U.S.: 13% of Americans speak Spanish at home (Dietrich & Hernandez, 2022), and it remains the most popular second language (American Academy of Arts & Sciences,

2016). Thus, substantial portion of the American population may prefer reading poetry in Spanish. However, since English is the primary language of *poets.org*, most works are translated into English, which may affect the representation of LXA poets in the collection. Nonetheless, increasing the number of LXA poems, whether in the original language or translated, would further enhance the collection's already strong diversity and equity. This enhancement is also essential for reducing potential biases in AI models for poetry caused by data imbalance.

Groupwise exclusive top word analysis

Figure 2 shows the top 10 unique words exclusive to each group to explore the word diversity and cultural depth each contributes. AA poems feature colloquial language reflective of African American Vernacular English (AAVE) (Khera, 2021), with terms such as 'lawd,' 'hyeah,' 'souf,' 'lovah,' and 'whah,' which mean 'Lord,' 'hear,' 'south,' 'lover,' and 'where,' respectively. APA-AA poems are rich in cultural references and names common in Asian contexts, such as 'Shiratama' (a Japanese dessert), 'Chang' (a family name in Korea or China), 'lola' (grandmother in Tagalog), as well as names such as 'Acequia.' APA-PA poems prominently feature words from indigenous languages, including Hawaiian Pidgin (Roberts, 1995), such as 'nalani,' 'kai,' 'huki,' 'olelo,' meaning 'the heavens', 'sea', 'to pull,' 'language,' respectively, along with names of places they live, such as 'guam' and 'hawai.' LXA poems contain many Spanish words, including 'tata,' 'une,' 'alabanza,' and 'templo,' which translate to 'grandfather,' 'article,' 'praise,' and 'temple.' Finally, NA poems incorporate historically significant words that describe their tribes or towns, such as 'Spavinaw,' 'Shawnee,' and 'Anishinaabeg,' as well as terms like 'chieftain' and 'clans' to depict their societal organization. The inclusion of non-standard English and foreign words underscores the diversity that these groups bring to American poetry. However, this diversity also highlights the potential limitations of linguistic analysis tools, including named entity recognition when applied to American poetry, especially if they were not trained from non-standard English varieties such as Hawaiian Pidgin or African American Vernacular English, or on foreign terms and languages.



Figure 2. Word clusters of individual groups

Groupwise popular theme analysis

To examine the similarities and differences among groups, we first identified each group's top 10 most popular themes. To further assess the thematic diversity that each group contributes to the collection in comparison with general poems, we first subtracted the theme proportions found in the general poems from those in each specific group. We then ranked these differences and selected the top 10 themes that were most distinctively prevalent in each group. Table 1 presents these themes and their corresponding percentages: the numbers in parentheses in the first five columns represent the proportion of each theme within a group, while the numbers in the last four columns show the difference from the corresponding proportions in general poems.

The dominant themes in general poems such as nature, love, death, body, and existential, illustrate a broad range of human emotions and experiences. Each underrepresented group has its distinct set of most popular themes, which only partially overlap with those of general poems, indicating their unique thematic focuses. Specifically, the NA group has only two overlapping themes, LXA has three, while AA and APA each have six overlapping themes. Furthermore, we examined the relatively popular themes in comparison to the general poems to understand the unique thematic focuses within individual groups. Among all groups or three out of the four, themes such as America, ancestry, body, identity, and family are prevalent. History, immigration, migration, and

social justice are also prominent, appearing in two groups. These common themes among them reveal how their experiences as underrepresented groups in America impact their poetry. Unique themes present in only one group include beauty, death, and slavery in AA; fathers, mothers, and politics in APA; memories and violence in LXA; and earth, environment, landscapes, language, and nature in NA. These highlights unique cultural and historical characteristics of each group. Particularly, NA poetry stands out with the most distinctive themes, reflecting their unique history as the original inhabitants prior to European arrival, and their strong connection to nature and environment. The distinct sets of popular themes among underrepresented groups raise concerns regarding the current theme classification systems. As theme sets are typically selected based on average theme popularity, as shown in Choi (2023), themes that are popular only within underrepresented groups often get overlooked. To create more equitable AI systems for poetry, we suggest developing multiple sets of classifiers for underrepresented groups, not only generalized but also tailored to their unique themes.

	Top 10 Most Popular Themes					Top 10 Most Relatively Popular Themes			
	General	AA	APA	LXA	NA	AA	APA	LXA	NA
1	nature (9.6)	identity (15.2)	body (11.6)	body (14.3)	nature (22.1)	identity (11.0)	identity (6.8)	identity (9.4)	ancestry (15.9)
2	love (8.8)	body (12.3)	identity (10.9)	identity (13.6)	ancestry (17.8)	america (9.0)	family (6.0)	immigration (8.0)	nature (12.5)
3	death (8.4)	america (12.1)	family (9.3)	death (10.8)	body (13.4)	ancestry (7.6)	ancestry (5.9)	family (7.5)	environment (9.7)
4	body (7.9)	death (11.9)	death (9.0)	family (10.8)	landscapes (11.9)	history (5.6)	immigration (5.4)	america (7.0)	earth (8.6)
5	existential (7.6)	love (10.8)	nature (8.9)	america (10.1)	environment (11.9)	social justice (5.4)	america (4.1)	ancestry (6.4)	america (8.4)
6	self (6.2)	ancestry (9.4)	self (8.6)	immigration (8.5)	america (11.5)	body (4.4)	body (3.7)	body (6.4)	landscapes (7.7)
7	beauty (5.6)	self (9.1)	existential (7.7)	ancestry (8.3)	history (10.7)	death (3.5)	mothers (3.7)	social justice (4.6)	family (7.4)
8	animals (5.3)	nature (9.2)	ancestry (7.7)	memories (7.8)	family (10.7)	beauty (3.3)	migration (3.1)	violence (4.5)	language (7.3)
9	loss (5.2)	beauty (9.0)	america (7.2)	violence (7.3)	earth (10.7)	slavery (3.3)	politics (3.1)	memories (3.8)	history (7.2)
10	writing (4.5)	history (9.0)	animals (6.7)	loss (7.0)	language (10.3)	hope (3.1)	fathers (3.1)	migration (3.5)	body (5.5)

Table 1. Top 10 most popular themes per group and relatively popular themes per group against general

Entropy-based analysis of theme distribution diversity

We explore the theme diversity within each individual group of poets to assess their contribution to the collection's overall diversity. For this analysis, we use entropy, given that this has been widely used to assess the diversity of systems and environments based on the richness and evenness of values (Jost, 2006). Entropy is defined as $H = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$ where H is the entropy, n is the number of themes, $p(x_i)$ is the proportion of the i-th theme within the poem group, and the summation is across all themes A higher entropy suggests a greater number of associated themes per poem (richness) and/or a more equitable distribution among them (evenness).

Figure 3 displays the cumulative distribution functions (CDF) of all groups, which represents evenness of theme distribution. The x-axis represents the proportion of themes within each group, while the y-axis represents the cumulative probability. A curve that rises quickly to 1 indicates low evenness, as this suggests that a few themes dominate; the corresponding histogram would show a steep decline. Conversely, a curve that rises slowly and stops at a larger theme proportion indicates high evenness, suggesting a more uniform distribution of themes; the corresponding

histogram would decrease more gradually. Besides the CDF, the associated entropy scores and the average number of themes associated with each group, as a measure of richness, are presented in the legend of the figure. Because APA-PA associates with fewer than half of the themes, it cannot be compared with others fairly. So, we merged it back to the APA category instead of distinguishing it from Asian Americans in this graph.

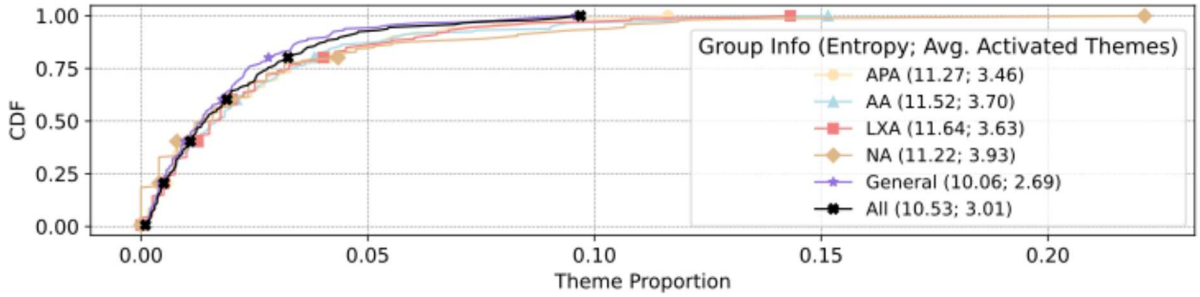


Figure 3. Cumulative distribution function (CDF) of theme proportions across groups, with corresponding entropy scores and average number of associated themes per group

The analysis of entropy scores shows a consistent trend: underrepresented groups achieve higher scores than their general counterparts, with scores of 10.06 for General, 11.27 for APA, and 11.64 for LXA. Furthermore, incorporating the underrepresented groups' poems into the general collection has increased the overall diversity by 0.47. To assess the contribution of richness to the entropy scores, we examined the average number of associated themes in each group. Compared to the General category's 2.69, poems from underrepresented groups tend to have more themes annotated, ranging between 0.77 and 1.24 more themes, which supports *poets.org*'s extra care toward these groups. While there is a correlation between entropy scores and thematic richness, this increase is also attributed to the evenness observed in the CDF in Figure 3. The CDFs of the underrepresented groups rise more gradually and stop at a larger theme proportion compared to the CDF for general poems. Similarly, although subtle, the CDF for all poems also ascends more slowly and stops at a slightly higher theme proportion, indicating an evenness contributed by underrepresented groups. Overall, our findings indicate that poems by underrepresented groups contribute to the collection's higher thematic diversity, benefiting from *poets.org*'s more thorough annotation of these poems and their more even distribution of theme proportions.

Conclusion

This paper evaluates *poets.org*'s poetry collection for racial and ethnic representation, addressing a gap in previous AI and NLP studies on poetry that overlooked the collection's assessment, potentially introducing biases against underrepresented groups. The collection generally reflects the demographics of the US population, with most categories having higher representation, except for LXA poems. The groupwise word and theme analyses show the diversity these groups bring to the collection, stemming from their culture and history. However, AI systems need to accommodate non-standard English and foreign terms to be more inclusive. Moreover, tailored theme sets for each group, rather than relying on those for the average, could ensure more equitable and inclusive AI systems for poetry. This study focuses on racial and ethnic diversity, one of many facets of diversity. In future work, we will further investigate other aspects, such as gender and disability.

Acknowledgments

This work was supported by RE-252382-OLS-22 from the institute of museum and library.

About the authors

Kahyun Choi is an Assistant Professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign. She earned her Ph.D. from the School of Information Sciences at the University of Illinois at Urbana-Champaign. Kahyun Choi's research interests involve the application of computational methods and machine learning algorithms to various modalities, including audio and text. She can be contacted at kahyun@illinois.edu

Gyuri Kang is a PhD student in Information Science at Indiana University Bloomington. Her research interests include digital environmental humanities, cultural analytics, and NLP. She can be contacted at gyukang@iu.edu

References

American Academy of Arts & Sciences. (2016). The state of languages in the U.S.: A statistical portrait. Cambridge, MA: Commission on Language Learning, American Academy of Arts & Sciences.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Choi, K. (2023). Computational thematic analysis of poetry via bimodal large language models. Proceedings of the Association for Information Science and Technology, 60(1), 538-542. https://doi.org/10.1002/pra2.812

Cordell, R. (2020). Machine learning and libraries: a report on the state of the field. Library of Congress.

D'ignazio, C., & Klein, L. F. (2023). Data feminism. MIT press.

Dietrich, S., & Hernandez, E. (2022). Language use in the United States: 2019. American Community Survey Reports. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2022/acs/acs-50.pdf

Drager, K. (2012). Pidgin and Hawai 'i English: an overview. International Journal of Language, Translation and Intercultural Communication, 1, 61-73.

Iyengar, S. (2023). New Survey Reports Size of Poetry's Audience, Streaming Included. Accessed: April 7, 2024.

Iyengar, S., Nichols, B., Shaffer, P.M., Menzer, M., Grantham, E., Santoro, H., Moyseowicz, A., & Hall, E. (2018). US trends in arts attendance and literary reading: 2002–2017.

Jo, E. S., & Gebru, T. (2020, January). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 306-316). https://doi.org/10.1145/3351095.3372829

Jost, L. (2006). Entropy and diversity. Oikos, 113(2), 363-375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Kaur, J., & Saini, J. R. (2017, February). Punjabi poetry classification: the test of 10 machine learning algorithms. In Proceedings of the 9th international conference on machine learning and computing (pp. 1-5). https://doi.org/10.1145/3055635.3056589

Khera, T. (2021). What Makes African American Vernacular English Distinct and Complex. Dictionary. com. Dictionary. com, February 21.

Lou, A., Inkpen, D., & Tanasescu, C. (2015). Multilabel subject-based classification of poetry. Nature, 2218, 30-7.

Navarro-Colorado, B. (2018). On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry. Frontiers in Digital Humanities, 5, 15. https://doi.org/10.3389/fdigh.2018.00015

Rakshit, G., Ghosh, A., Bhattacharyya, P., & Haffari, G. (2015, December). Automated analysis of Bangla poetry for classification and poet identification. In Proceedings of the 12th international conference on natural language processing (pp. 247-253).

Stevenson, D. (2021). Application of Shannon Entropy Metrics to Cultural Diversity and Language Evolution. Academia Letters, 2.

U.S. Census Bureau. (2022). Race and Hispanic origin. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045222

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Towards memetic legitimation of knowledge: memes and cultural heritage

Alexander O. Smith, Calvin Cousin, Una Joh, Christy Khoury, Yiran Duan, and Jeff Hemsley

DOI: https://doi.org/10.47989/ir30iConf47260

Abstract

Introduction. We introduce legitimation practices through Know Your Meme documenting the high-profile sex-trafficking and suicide ruling of Jeffrey Epstein.

Method. Using netnography, we analyse legitimation discourse provided by the six authors reported previous knowledge and what Know Your Meme legitimised about memetic knowledge, the website, and the Epstein case and surrounding conspiracy.

Analysis. Our preliminary analysis suggests that legitimation discourse is a valuable netnographic methodology to understand how documentation both technically and textually establishes '*authoritative*' contexts for memetic and conspiracy theoretic content. We have found that Van Leeuwen's categories of legitimation function in relation to self-reported claims about cultural heritage sites such as Know Your Meme's archival and narrative accounts of memetic meaning.

Conclusion. Our work begins filling two gaps in contemporary digital information environments. Firstly, we begin to close the gap between netnography and cultural heritage sites such as archives, repositories, and so forth. Secondly, we contribute diverse individual legitimations of memetic content. Future work will develop more rich interpretations of what is legitimated by Know Your Meme. Additionally, we compare individual participants' legitimation process relative to their prior reported knowledge.

Introduction

Conspiracy theories are abundant online, sometimes hiding within otherwise verifiable claims. Whether '*true*' or not, conspiracy theories exemplify cultural processes. Especially through social media and documentation practices, they gain legitimacy in popular discourse. As such, conspiracy theories exemplify *cultural heritage* understood as a media imbued with the power of memory (Taylor, 1995; Viejo-Rose, 2015).

A recent example of media and cultural memory surrounds Jeffrey Epstein's sex-trafficking case and the ambiguity around who was involved and what occurred. Epstein was a New York financier who was arrested and convicted for sex-trafficking charges in 2019. A part of the motivation for the conspiracy around Epstein was the legitimation provided by the *documented* connections to social elites. Associating with powerful figures like the Clintons and Donald Trump generated speculation about the involvement of social elites in Epstein's human trafficking (Colyar et al., 2019). Refuelling interest, Epstein's travel logs to and from his private island – where much of the sex trafficking took place – was released in January of 2024 including many high-profile names. However, the legitimacy of the narrative is elusive. Despite his death being ruled a suicide, there exists competing documented evidence about Epstein's death through autopsy reports. Additionally, motivation and opportunity existed for any co-conspiring social elites to have him murdered. In response, the '*Epstein didn't kill himself*' meme/conspiracy theory circulated widely (Stockler, 2019).

Like this meme, conspiracy theories are linguistic claims about political knowledge that are often difficult to validate (Zannettou et al., 2019). Yet, the declarative statement, '*Epstein didn't kill himself,*' is unusual for memes. Memes often provide contexts that resist immediate linguistic reduction (Yus, 2019). Regardless, memes often elusively shape worldviews (Wiggins, 2019) much like conspiracy theories. We make use of Know Your Meme (KYM) entries that document and legitimise perspectives surrounding Epstein conspiracy theories. A meme serves as affiliated digital media traces that shape and sustain knowledge as cultural heritage.

This paper contributes to cultural heritage research by making use of KYM's perceived authority in generating connections between memetic contexts and personal memory and knowledge. It does so by using the authors' varying degrees of knowledge surrounding Epstein, and the knowledge/perspectives the KYM facilitated for us. We further used netnography to analyse how memetic content legitimises aspects of the Epstein case. Using legitimation discourse (Van Leeuwen, 2007) from the authors' produced texts, we respond to the following questions:

> RQ1. How do memes surrounding Jeffrey Epstein legitimate people's perspectives of knowledge surrounding conspiracy?
>
> RQ2. Does KYM generate or instigate memetic knowledge that is incomplete, fabricated, or limited?

The first question is about reinforcement/disjunction of what was remembered by the researchers. The second question focuses on the idea that perhaps some of what KYM does is afford systematic documentation of cultural events. This question is designed to ask about how meme data as organised by KYM might privilege a particular kind of cultural heritage.

Literature review

KYM is a website that contextualises *internet memes*: collections of artefacts that spread online, carrying some common '*content*, *form*, *and/or stance*' (Shifman, 2014, p. 41). KYM provides documented references for internet memes and digital culture. The site's curators validate meme entries developed through crowdsourced labour of artefact collections and metadata. With few exceptions (e.g., Bratt & Smith, 2022; Pettis, 2021; Rogers & Giorgi, 2023) KYM is under-studied.

KYM curates content based on the associative quality of cultural context, contributing to the understanding of memes' meanings. KYM attempts to account for these meanings with ambiguous degrees of completion (Pettis, 2021). The platform provides histories of memes, accounting for the way people position or perform in relation to memetic content. As such, KYM is an exemplary cultural heritage site of memes. Memes and how they are documented as cultural heritage is a particularly recent research interest (Bratt & Smith, 2022; Modrow & Youngman, 2023; Rees, 2021; Tulloch, 2023), playing an increasing role in more recent information studies' contributions to memetics (see, Smith et al., 2024).

KYM's curation provides new interpretations of culture through technological organisation (Rogers & Giorgi, 2023) and internet history (Pettis, 2021). As such, KYM generates memetic '*authority*' for possible truth claims about the memetic content. KYM is a site that focuses on internet culture by sorting piles of digital ephemera using crowdsourced labor. The site has a '*roguish*' (Kosnik, 2021) and '*dumpster-esque*' (Hogan, 2015) quality to its piles of content, where the archival meaning is inductively assembled. As such, in aggregation, KYM provides a post-humanesque generation of living media and memory (Bennett, 2010). These roguish collections of ephemera are granular, and not necessarily narrative in their parts. Yet, they coalesce into material traces of memory (Smith & Loewen-Colón, 2024), suggesting legitimation from what otherwise might be disconnected without the archive. Using the memetic quality of the Epstein case limits the validated empirical knowledge accessible to researchers and forces a focus on the conspiratorial narrative provided by these post-human memetic aggregations. Nevertheless, the results of aggregating these ephemera into a dumpster-like collection requires questioning the human-focused narratives surrounding how individual people play a part in legitimating a conspiracy theory. As such, memes contribute exemplary tensions in documentative agency and power of archives, crowdsourced curators, and sourced actors in '*canonizing*' stakeholders into a morally charged conspiracy narrative (Mackinnon, 2022).

Prior discussion exists on the importance of memes in contextualising conspiracies (Hagen, 2022; Harvey et al., 2019; Wiggins, 2021b, 2021a). However, little work explains how memes *actively* provide context. We contribute a development of how memetic contexts integrate into legitimised knowledge of conspiracy theory. Additionally, we provide a methodological approach to cultural heritage generated through memes.

In developing notions of knowledge through memes, Harbo (2022) analysed (de)legitimation discourse practices of a Facebook group, considering memetic artefacts and related evaluative knowledge claims about economic theory. We use a framework of narrative knowledge: legitimation discourse analysis (Van Leeuwen, 2007). Below, provides our protocol for integrating this framework of legitimation into our netnographic data.

Methodology

Netnography (Kozinets, 2019) has diverse methodological applications in memetic research. Previous literature in memetics has used netnographic inspirations, specifically spanning topics like extremist ideological spread (Trillò & Shifman, 2021), adoption of religious ideas (Rohmatulloh, 2019), audience recontextualization of fictional narratives (Rendell, 2019), and memories and visual journaling (Terzimehić et al., 2021). A netnographic approach specifically for memetic content has also been developed (Kaltenhauser et al., 2021). In our specific case, netnography enables us to select a subset of linked KYM entries as a cultural heritage site, and have our author-participants use 'cultural engagement strategies' (Kozinets, 2019, p. 250) with the site. Netnography enables us to look at the documented symbolic developments related to the site's categories of legitimacy (Van Leeuwen, 2007) relative to our diverse interactions with meme content.

In our protocol, we netnographically (Kozinets, 2019) evaluated eight meme entries' descriptions and image galleries connected to Jeffrey Epstein. Netnography is frequently a method used to

evaluate the discursive qualities of memes and surrounding discussions (e.g., Nissenbaum & Shifman, 2017). Kozinets' outline of netnography is particularly useful for this case for several reasons. Firstly, it allows us to look at the pragmatic cultural implications of digital media's interpretation (Kozinets, 2019, pp. 16–17). Also as KYM is both a curatory site and a memetically investigatory resource, it fills a gap by bridging the digital curatory aspects that Kozinets (2019, pp. 95–96) notices is understudied from the netnographic lens with the digital traces and the suggested external knowledge to those witnessing these traces. In particular, netnography enables us to interpret KYM's data as an ambiguous cultural reality and integrate it into our own personal knowledge. In turn, netnography provides investigatory tools to interpret the knowledge provided by this memetic reality as a culturally situated integration of trace data as newly lived memories of cultural heritage.

Within our netnography, we designed questions informed by Van Leeuwen's (2007) framework, motivating researchers to reflect on gained knowledge. Van Leeuwen provides four categories of legitimation: authorization, moral evaluation, rationalisation, and mythopoesis. *Authorization* is how authority is discussed. The netnography of KYM suggests how authority is given to the memetic context of artefacts and the textual analysis of entries for each meme entry on the website. *Moral evaluation* considers how value systems are implied concerning the entries. The netnographic values are extracted surrounding how researchers moralise the claims KYM provides to them. *Rationalisation* is how knowledge is justified. We evaluate netnographic rationalisations of the knowledge and moral claims. *Mythopoesis* is how knowledge is narrated. We evaluate mythopoesis by considering how our researchers narrate their own knowledge through the site's curated information.

The first two authors designed a set of questions for each researcher to report their knowledge about memes, KYM, and the Epstein case/conspiracy. Then, questions were designed for all the authors to interpret the meme entries that prompted researchers to analyse (de)legitimation discourse surrounding KYM and how the information might be memetically '*authoritative*.' Together, the questions provided room for the researchers to narratively explain the legitimacy of the memetic content.
Each author evaluated a '*main*' entry, called '*Jeffrey Epstein sex trafficking case*', and one or two other entries related to the main entry using a Chrome-based browser. The 'main' entry is the primary KYM entry as it exists at the top of a hyperlink hierarchy. The other entries considered the conspiracy theories surrounding Epstein's suicide, his client list, the '*Epstein files*,' with information regarding powerful figures; his mansion; some of his connections, such as socialite Ghislaine Maxwell, politician Donald Trump, and British royal Prince Andrew.

We independently followed a set of structured survey questions and instructions on how to answer them. Additionally, our authors Zoom-record their process of analysis. Some of the questions asked about the entries were as follows: How did the entry frame Jeffrey Epstein? Who are the characters involved and what is their relation to the case? How do the visual elements complement the encyclopaedic entries? How can the artefacts be interpreted as a *'meme'*?

In the next session, using the legitimation categories of Van Leeuwen (2007), we analyse how the researchers viewed the content of their assigned entries and how said content suggested knowledge about the Epstein case and his death.

Results and discussion

First, we introduce our authors in summary. It is of cultural significance to situate who the participants/authors are. Our background knowledge and diverse cultural contexts shape our understanding of memes and Epstein. For this, each author self-reported our knowledge of memes, KYM, and Epstein in a document prior to our analysis.

All six authors served as our participants, all of whom are PhD students or university faculty spanning two Universities in two countries. We range in age from mid-20s to late-40s, being male, female, agendered, and spanning many different ethnicities, carrying citizenship to at least four different countries. So, while we are all focused within a research university setting, otherwise we are a diverse research team. As such, our individual legitimation from memes, KYM, and Epstein is more robust, enabling us to have more diverse cultural contexts for approaching the formation of cultural heritage as documented on the site.

Of our authors, Smith and Hemlsey have published memetic research while the rest have not. Self-reported knowledge of memes ranged from more academic to casual social interaction. Some understood memes more as '*visual artefacts*' shared with their friends, while others reported academic conceptions of collections of artefacts (Shifman, 2014) or semiotic references (Cannizzaro, 2016). Most of the authors had heard of KYM or had interacted with the site briefly. Smith and Hemsley were very familiar with the site. Self-reported knowledge of Epstein ranged from detailed facts about the case to incorrect geographic information about Epstein's location. Duan developed a Twitter data collection surrounding Epstein before the study, but otherwise there was no academic involvement in Epstein or conspiracy theory literature.

Regarding the main entry, the researchers identified memes being used as part of the construction of the narrative. However, we experienced difficulty understanding *how* the meme entry was '*a meme.*' Instead, KYM was identified as being like an encyclopaedia, presenting data as factual. We noted that KYM editors seem to frame sympathy with the narrative that Epstein didn't kill himself, but they do not go as far as stating it was not suicide. Rather the editors provide us with hints for how Epstein might not have killed himself. While embodying the encyclopaedic genre, defending an editorial view with data, the information presentation appeared playful and humorous in tone. It also felt like an invitation to learn more about the situation, either by the hyper-textuality of the platform or through other means.

In our initial analysis, we see some initial evidence of legitimation practices surrounding the conspiracy of Epstein's case. Van Leeuwen's (2007) categories helped identify that *authorization* is used in the entries by using constant affirmative sentences; every entry states the events related to it as a fact, even if absurd. *Moral evaluation* may consist in the way the website presents the actions of Epstein and some of his clients as actual crimes, with some of the entries having warnings of sensitive content and framing that he might have not killed himself. The way the events are presented, inclusive of evidence for possible conspiracies, does not frame them as unimportant or morally correct. Regarding *rationalisation*, every entry presents sources for what is claimed, either official, from news outlets or social media posts that contribute to what is being presented. For the *mythopoesis* provided by the entries, our researchers reported they appear as encyclopaedic pages: there is an introduction of the event and a further description of it, with links leading to related entries and to outside sources that may give to the entry its dispersion characteristics. Every entry also has a gallery of the images and videos of it. Overall, it is identified that the content of the entries makes sense within their own logic.

The secondary meme entries mostly focused on conspiracy theories themselves or powerful figures connected to Epstein (e.g., Donald Trump, Prince Andrew, Ghislaine Maxwell). These entries mostly reinforced the researchers' understanding of the main entry. The frequent use of humour in entries and artefacts act to motivate through pathos via '*meme*,' engaging the reader and contributing to the rationalisation of what is being presented, the mythopoetic way in which those entries are narrated. The visual artefacts manage to provide a greater understanding of the contexts, acting as evidence of what is being said, further contributing to the rationalisation of the entries. The visual elements not only added to the context but were themselves often described in detail in the entries. More than one researcher noted that reading the entries enticed them to learn more about the subject, either to have a larger understanding or simple curiosity about the

individuals involved. Thus, in the presentation of various events as facts, including the dubious ones, and using an online encyclopaedic format of various memes to come up with a larger story, we could say that KYM does contribute to the generation and instigation of memetic knowledge, even if what is being presented is fabricated or incomplete. By using an authoritative tone and coming up with multiple sources to back up its statements, KYM shapes the narratives it presents as both memes and factual events. Its uses of mythopoesis and rationalisation, in particular, contribute to the feeling of legitimation of the entries.

Closing remarks

We have shown promise in filling two gaps: one in netnographic approaches to cultural heritage sites and a second in internet memetics. First, our work begins filling a gap between netnography and cultural heritage sites, like archives. As noted by Kozinets (2019), application of netnography has been scarce. However, our use of legitimation discourse specifically responds to how individuals report gained knowledge netnographically. Second, we contribute diverse individual legitimations of memetic content, moving toward understanding how unique and diverse individuals recall pre-linguistic memetic content's meaning within their own words.

The study of this one narrative works as an initial study on how legitimation emerges from KYM, although it is not an exhaustive investigation of the many ways this may happen. Future work will develop richer interpretations of what is legitimated by the platform. In doing so, we aim to develop a more full understanding of the narrative power dynamics extending legitimation to an '*archival promenade*' (Mackinnon, 2022) ethics-oriented methodology, considering all actors, visible and invisible. We will compare individual participants' legitimation process relative to their prior reported knowledge. Our netnographic writing suggests that the application of Van Leeuwen's legitimation practices can be more fully evaluated through his categories to contribute a methodological approach more fully for cultural heritage sites. Further, we notice the potential for unique '*knowledge*' about the conspiracies as seen through subjective paths through data, through a variety of documented entries.

Acknowledgements

We would like to thank the reviewers for their time and effort in evaluating our work.

About the authors

Alexander O. Smith is a PhD candidate of Information Science and Technology at Syracuse University. Their research is focused on memetics: the study of memes. Alexander is forwarding a novel theoretical framework for memetics which synthesizes older memetic theories with internet meme studies. Affiliated work can be found at *Convergence*, *Journal of Documentation*, *ASIS&T*, and the *IEEE Big Data Conference* Computational Archive Science workshop. More expansive research interests on (cross-)platform dynamics, network cultures, and classification can be found at *policy and internet*, *social media + society*, *eCSCW*, and *AoIR*. They can be contacted at aosmith@syr.edu

Calvin Cousin is a PhD candidate at the College of Library Sciences and Communication at Federal University of Rio Grande do Sul (UFRGS, Brazil) and a visiting scholar at Syracuse University. His research interests include digital influencers, disinformation, and online discourse. Currently, his main research focuses on the role of celebrities in online discussions about disinformation during the 2022 Brazilian presidential elections. Calvin has a background in journalism, having taught convergent journalism and communication and museums courses at UFRGS, and is a member of the Association of Internet Researchers (AoIR). He can be contacted at calvin.cousin96@gmail.com

Una Joh is a Ph.D. student at the School of Information Studies at Syracuse University, specializing in computational social science. With a background in economics from her bachelor's and master's

studies at SKKU in Korea, she conducts research on the flow of information and the dynamics of public opinion on social media, taking into account its interaction with news media. Her research aims to extract, track, and predict information flows and opinion dynamics on networks that model interconnected media environments. To achieve this, she utilizes a variety of methods, including natural language processing, network analysis, and agent-based modeling. She can be contacted at sjoh01@syr.edu

Christy J. Khoury is a Ph.D. Candidate at the Syracuse University iSchool and a 2024 Research Excellence Doctoral Fellow. Christy's research, situated in the crisis informatics space, uses a sociotechnical perspective to examine the uses of TikTok for sensemaking of and resiliency to crises that occurred in Lebanon. Her dissertation uses a mixed-methods approach to explore TikTok as a contemporary crisis communication tool. She has also researched Reddit, Twitter, Facebook, and Instagram to examine discourse centered around crisis and non-crisis contexts with the Network Influence and Information Curation (NIIC) lab. Overall, her research takes on a user-informed approach to consider the interaction between users and technology in a time of uncertainty. Her research endeavours are rooted in a transnational, information ethics lens that mirrors her pedagogical perspectives. She can be contacted at cjkhoury@syr.edu

Yiran Duan is a PhD candidate at Syracuse University's School of Information Studies. Her current research interests are at the intersection of cross-platform comparison and social network analysis through the lens of curated flow. Specifically, she explores how information travels across platforms as well as the lifespan of information flows. She approaches her research topics by applying both qualitative and quantitative methods such as content analysis and machine learning models. Her work has been published in the *Association of Information Science and Technology* (ASIS&T), the *iConference*, the *Association of Internet Researchers* (AoIR), and *social media + Society*. She can be contacted at yduan12@syr.edu

Jeff Hemsley is the Interim Dean at the School of Information Studies at Syracuse University. He earned his Ph.D. from the University of Washington's Information School. His research is about understanding information diffusion, particularly in the context of politics, in social media. He is co-author of the book Going Viral (Polity Press, 2013 and winner of ASIS&T Best Science Books of 2014 Information award and selected by Choice magazine as an Outstanding Academic Title iConference for 2014), which explains what virality is, how it works technologically and socially, and draws out the implications of this process for social change. He can be contacted at jjhemsle@syr.edu

References

Bennett, J. (2010). Vibrant Matter: A Political Ecology of Things (1st ed.). Duke University Press Books.

Bratt, S., & Smith, A. O. (2022). Evolutionary Archives: The Unlikely Comparison of GenBank and Know Your Meme. 2022 IEEE International Conference on Big Data (Big Data), 2490–2499. https://doi.org/10.1109/BigData55660.2022.10020929

Cannizzaro, S. (2016). Internet memes as internet signs: A semiotic view of digital culture. Internetimeemid Kui Internetimärgid: Semiootiline Pilk Digitaalkultuurile., 44(4), 562–586. https://doi.org/10/gf8wfk

Colyar, B., Hurwitz, K., Klein, C., Kweku, E., Larocca, A., Martins, Y., Raymond, A. K., Schneier, M., Stieb, M., & Walsh, J. D. (2019, July 22). The High Society That Surrounded Jeffrey Epstein.

Intelligencer - New York Magazine. https://nymag.com/intelligencer/article/jeffrey-epstein-high-society-contacts.html

Hagen, S. (2022). 'Who is /ourguy/?': Tracing panoramic memes to study the collectivity of 4chan/pol/. New Media & Society, 14614448221078274. https://doi.org/10/gp3s5s

Harbo, T. F. (2022). Internet memes as knowledge practice in social movements: Rethinking Economics' delegitimization of economists. Discourse, Context & Media, 50, 100650. https://doi.org/10.1016/j.dcm.2022.100650

Harvey, A. M., Thompson, S., Lac, A., & Coolidge, F. L. (2019). Fear and Derision: A Quantitative Content Analysis of Provaccine and Antivaccine Internet Memes. Health Education & Behavior, 1090198119866886. https://doi.org/10/gf8whz

Hogan, M. (2015). The Archive as Dumpster. Pivot: A Journal of Interdisciplinary Studies and Thought, 4(1), Article 1. https://doi.org/10.25071/2369-7326.39565

Kaltenhauser, A., Terzimehić, N., & Butz, A. (2021). MEMEography: Understanding Users Through Internet Memes. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 1–7. https://doi.org/10/gj3jhs

Kosnik, A. D. (2021). Rogue Archives: Digital Cultural Memory and Media Fandom. MIT Press.

Kozinets, R. V. (2019). Netnography: The Essential Guide to Qualitative Social Media Research. SAGE.

Mackinnon, K. (2022). Critical care for the early web: Ethical digital methods for archived youth data. Journal of Information, Communication and Ethics in Society, 20(3), 349–361. https://doi.org/10.1108/JICES-12-2021-0125

Modrow, S., & Youngman, T. (2023). Theorizing Cultural Heritage Informatics as the Intersection of Heritage, Memory, and Information. Proceedings of the Association for Information Science and Technology, 60(1), 666–671. https://doi.org/10.1002/pra2.836

Nissenbaum, A., & Shifman, L. (2017). Internet memes as contested cultural capital: The case of 4chan's /b/ board. New Media & Society, 19(4), 483–501. https://doi.org/10/f99h7v

Pettis, B. T. (2021). Know Your Meme and the homogenization of web history. Internet Histories, 0(0), 1–17. https://doi.org/10/gmkh47

Rees, A. J. (2021). Collecting Online Memetic Cultures: How tho. Museum and Society, 19(2), 199–219. https://doi.org/10/gmfdcw

Rendell, J. (2019). A picture is worth a thousand corpses: Audiences' affective engagement within the Flesh and The Walking Dead through online image practices.

Rogers, R., & Giorgi, G. (2023). What is a meme, technically speaking? Information, Communication & Society, 0(0), 1–19. https://doi.org/10.1080/1369118X.2023.2174790

Rohmatulloh, D. M. (2019). In Meme Dakwah: A Netnographic Study of Garis Lucu Social Media Accounts. The 19th Annual International Conference on Islamic Studies (AICIS 2019). https://doi.org/10.2139/ssrn.4042250

Shifman, L. (2014). Memes in Digital Culture. The MIT Press.

Smith, A. O., Hemsley, J., & Tacheva, Z. Y. (2024). Reinforming memes: A literature review of the status of memetic information. Journal of Documentation, 80(4), 1003–1021. https://doi.org/10.1108/JD-10-2023-0218

Smith, A. O., & Loewen-Colón, J. (2024). Memetic memory as vital conduits of troublemakers in digital culture. Convergence. https://doi.org/10.1177/13548565241262421

Stockler, A. (2019, August 27). Epstein Lawyers Tell Judge Murder Is More Likely Explanation of Death. Newsweek. https://www.newsweek.com/epstein-death-murder-suicide-victims-hearing-1456397

Taylor, H. A. (1995). "Heritage" Revisited: Documents as Artifacts in the Context of Museums and Material Culture. Archivaria, 8–19.

Terzimehić, N., Schött, S. Y., Bemmann, F., & Buschek, D. (2021). MEMEories: Internet Memes as Means for Daily Journaling. Designing Interactive Systems Conference 2021, 538–548. https://doi.org/10.1145/3461778.3462080

Trillò, T., & Shifman, L. (2021). Memetic commemorations: Remixing far-right values in digital spheres. Information, Communication & Society, 0(0), 1–20. https://doi.org/10/gm7jd6

Tulloch, B. J. (2023). Memes to an End: Why Internet Memes Matter to Information Research. Libri, 73(4), 293–305. https://doi.org/10.1515/libri-2021-0098

Van Leeuwen, T. (2007). Legitimation in discourse and communication. Discourse & Communication, 1(1), 91–112. https://doi.org/10.1177/1750481307071986

Viejo-Rose, D. (2015). Cultural heritage and memory: Untangling the ties that bind. Culture & History Digital Journal, 4(2), e018. https://doi.org/10.3989/chdj.2015.018

Wiggins, B. E. (2019). The discursive power of memes in digital culture: Ideology, semiotics, and intertextuality. Routledge.

Wiggins, B. E. (2021a). Boogaloo and Civil War 2: Memetic antagonism in expressions of covert activism. New Media & Society, 23(11), 3179–3205. https://doi.org/10.1177/1461444820945317

Wiggins, B. E. (2021b). #Civilwar2: Instagram posts during COVID-19. Social Semiotics, 31(3), 402–420. https://doi.org/10/gkq8rx

Yus, F. (2019). Multimodality in memes: A cyberpragmatic approach. In Analyzing digital discourse (pp. 105–131). Springer.

Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. Journal of Data and Information Quality, 11(3), 1–37. https://doi.org/10/gghb7s

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Innovation hesitancy: exploring reluctance to digital innovation in the Swedish cultural sector

Hanna Carlsson, Ahmad Kamal, and Charlie Järpvall

DOI: https://doi.org/10.47989/ir30iConf47245

Abstract

Introduction. This paper explores a challenge for any novel digital intervention: the hesitancy or reluctance among the designated user group.

Purpose. The overall purpose is to better understand the hesitant reception of digital innovations through a specific case study. The aim is to further our understanding of resistance towards digital innovation as well as reflect upon the pro-innovation bias of digital development projects.

Method. Multiple sources were combined: data collected within the case project and conducted document analysis of project documentation.

Results. Labor and maintenance were the main reasons behind stakeholder hesitancy, as well as concerns regarding the hosting, funding, and local relevance of the proposed innovation.

Conclusion. The assumption that digitization is inherently beneficial shaped the direction of the case project. Yet ongoing cultural changes and current adaptations must be understood to ensure a new digital solution is actually needed. As researchers we must pay enough attention to relevant local conditions by critically engaging with the innovation process of the project itself. This requires looking pass the pro-innovation bias.

Introduction

This paper explores a challenge for any novel digital intervention: the hesitancy or reluctance among the designated user group. The paper carries out this exploration by analysing feedback from preliminary studies in the course of developing a digital platform aimed at the cultural sector in Sweden.

Since 2023, a Swedish co-creation research initiative set out to explore information sharing practices of cultural sector actors in Småland, a region in south-eastern Sweden. The aim of the project was to develop and introduce innovative tools and methods for collaboration between various cultural sector actors. One of the initiative's six subprojects – referred to hereafter as '*Project X*' - was devoted to digital innovation, with a focus on developing a shared platform, equipped with functions for searching for, discovering, and sharing information about cultural events across the region (and possibly beyond). This was in response to the lack of such a service in Småland, with information regarding local cultural events (performances, exhibits, demonstrations, public talks, fairs, educational events, special showings, etc.) being dispersed across various institutional sites, municipal event calendars, commercial marketing services, and other local sources (e.g., newspaper listings). Furthermore, end-users (the general public) were underserved by the limited functionalities of the various platforms carrying the relevant information, such as of the municipal calendars and social media sites used to publicize cultural events.

Project X identified this disaggregation as a missed opportunity to inform end-users about cultural events, potentially resulting in their non-participation. Such outcomes could have negative implications not only for cultural event providers but also for the broader local community. Cultural events, after all, play a crucial role in enhancing the attractiveness of a region, fostering social interactions, improving quality of life, and stimulating local economies. Additionally, the objective of ensuring that arts and culture are accessible to all as a public good is a central tenet of both national and local cultural policies in many European countries. Providing up-to-date and easily accessible information about cultural events to both residents and tourists is important for both economic development and the promotion of democratic engagement.

While some cultural sector actors expressed interest in the platform being developed within Project X, it was also received with a fair degree of hesitancy – at times elevated to disinterest or outright suspicion – by several cultural sector actors in Småland. In this paper we examine this hesitancy, applying critical innovation studies (e.g., Godin and Vink, 2017) and the concepts of pro-innovation bias and innovation resistance (cf. Fornstedt, 2021) to possibly explain these responses among the stakeholders as well as to critically reflect on the assumptions driving innovative interventions like Project X.

Purpose

The overall purpose of this study is to better understand the hesitant reception of digital innovations through a specific case study. Reporting on the initial findings of a larger research project, this paper examines innovation hesitancy expressed by actors who publicize cultural events. These actors included staff from cultural organizations (museums, libraries, archives, galleries, theatres, cultural festivals, cultural organizations, etc.) as well as regional coordinators (municipalities and their cultural departments/units). The object of study was the hesitant or reluctant attitudes of these actors expressed towards sharing information on local cultural events through a new digital platform. By critically engaging with these findings, our aim is to further our understanding of resistance towards digital innovation as well as reflect upon the pro-innovation bias that frequently seeps into digital development projects.

Perspectives on innovation

Resistance to and rejection of innovation have been studied from various perspectives. As pointed out by Fornstedt (2021), scholarly research on innovation resistance primarily focuses on identifying and understanding such aversions in order to assist managers and other stakeholders in mitigating or overcoming them (e.g., Stucki, 2019; Thomä, 2017). The bulk of this research focuses on private sector companies, commonly perceived as the primary drivers, incubators, and beneficiaries of innovation. The majority of this research adopts a pro-innovation perspective, with many studies linking the value of innovation to its capacity to enhance corporate survival and success (e.g., Heidenreich & Kraemer, 2016). Following this line of thought, there is also a prevailing tendency to categorize phenomena perceived as progressive under the term "innovation," while framing resistance to innovation as a reactionary opposition.

By contrast, critical innovation studies (CIS) questions the widespread tendency in contemporary society to understand and describe innovation as unequivocally positive. CIS critically interrogates how policymakers, businesspersons, researchers, and educators portray novelty and innovation as central components for virtuous societal improvement (Hallonsten, 2023; Pfofenhauer & Jasanoff, 2017). This perspective, referred to in CIS as the '*pro-innovation bias*', conceptualizes innovation as inherently and universally beneficial to the context in which it is introduced (Godin & Vinck, 2017). This value is frequently articulated in economic terms, which are assumed to yield broader benefits, such as the enhancement of people's quality of life. Consequently, solutions to various societal challenges are primarily sought in technological innovations; meanwhile the potential drawbacks or even harmful effects of innovation tend to be overlooked (Fougére & Harding, 2012; Fornstedt, 2021) and resistance – even within the scholarship – is problematized and deprecated (Fornstedt, 2021). Opponents of digital developments, for instance, are frequently dismissed as regressive or ignorant (cf. Wormbs, 2010). However, from a CIS perspective, resistance is instead viewed as a source of knowledge that enables a critical examination of the particular innovation at hand and of innovation more generally (Fornstedt, 2021). Hesitancy and resistance may even introduce the friction necessary for a more robust and inclusive approach to studying and implementing innovation.

This is particularly relevant with regards to the needs of marginalized groups whose voices are often neglected or even silenced in design and implementation processes of new innovations. This is often the case even in instances where they are directly affected by the outcomes of those processes (cf. Du, Xie & Waycott, 2020). CIS thereby seeks to move beyond a simplified narrative of innovation as quick fix to resolving structural power dynamics.

It is with this in mind that we turn the spotlight on the instances of hesitancy and rejection which greeted Project X's proposed innovation and how this hesitancy was framed by the cultural sector actors participating in the research.

Method

For the analysis we have combined multiple sources to explore the framing the innovation of Project X. We have analysed data collected within the project (essentially user studies) and conducted document analysis of project documentation, including the project application, plan, and description, along with notes from project meetings.

Insights on stakeholder hesitancy were gleaned from a survey, followed-up by a focus group interview. The initial step in Project X involved mapping relevant stakeholders within the cultural sector of Småland, encompassing both private and public entities, as well as those within the tourism industry. A total of 100 actors were identified through this process. In the second step, these actors were invited to complete an online survey consisting of 22 questions, which 34 respondents did. The survey solicited responses on various topics (e.g., marketing practices, digital events) as well as attitudes towards a new, wide-scale digital platform for advanced,

comprehensive, and up-to-date information on cultural events. To gain deeper insights into the survey responses, all participants in the survey were also invited to participate in a focus group. Although seven respondents accepted the invitation, only four attended. Participants were from the private cultural sector, tourism, and the public cultural sector. The interview lasted 40 minutes and was conducted via Zoom, with one researcher facilitating the discussion while the other took notes. No recordings were made due to restraints in the ethics approval. The emergent discussion shed further light on the basis of the innovation hesitancy observed in the survey.

For the analysis, central concepts from CIS were used in order to critically reflect upon the positions taken by the cultural sector actors participating in project.

Findings

Project documents

A review of the project application and plan for Project X reveals that its goals and motivations are strongly aligned with ideas and values characteristic of the pro-innovation bias. The idea of innovation as inherently and universally beneficial to the context in which it is introduced (e.g., Fougère et al., 2017) is illustrated in statements such as:

> *The project will demonstrate how to embrace digital change and encourage partnerships that foster innovation, also making it easier for people to use culture for education, research, creation, and recreation.*

Furthermore, the project attempts to provide solutions to societal challenges, in this case public access to culture, by means of technological innovations. Thus, potential drawbacks or even harmful effects of the suggested innovation are downplayed in the project documentation.

> *The project will therefore set out to invent new ways to collect, promote, and analyze information about cultural events, using relevant digital technologies to make cultural events more accessible to the public on the one hand while providing insights to event organizers on the other.*

One such harmful effect mentioned by the participants in both the survey and the focus group is the issue of labor and maintenance.

Stakeholder survey

The survey response rate was 34%, with half of the respondents (53%) representing municipalities while the rest were from diverse cultural organizations (e.g., museums, venues, galleries). Results showed that the most common platforms for publicizing cultural events were social media accounts, local websites, and municipality websites.

Platforms for publishing cultural events	
Social media accounts	97%
Local websites	82%
Municipality websites	70%

Table 1: Choice of platforms

When asked if respondents were willing to contribute information to a joint platform, approximately half of the respondents gave an affirmative. However, a nearly equal amount provided reasons for hesitancy or reluctance on their part, such as '*the need for more information about the platform*' or '*not having the resources to contribute*' (Interestingly, even a respondent

expressed unequivocal willingness to participate in the proposed platform gave reason hesitancy). One respondent succinctly stated: '*I don't believe in the idea.*'

Reasons for perhaps/no	Count
Need additional information to make a decision	10
Do not have the resources to contribute	7
Existing platforms are sufficient	1
Other: '*I don't believe in the idea*'	1

Table 2: Reasons for hesitancy

Focus group interview

The hesitancy and reluctance expressed in the survey was sizeable and a focus group interview allowed us to discursively explore this beyond the restricted format of an online questionnaire. Whereas the survey indicated interest among the respondents in a joint platform, the focus group participants exhibited more vehement reluctance. The participants expressed varying needs but unanimously conveyed little interest in a joint online platform for providing an overview of regional cultural events. They cited several reasons for this, the central ones being the related issues of timing and insufficiency of resources. Furthermore, participants revealed that – based on a failed previous attempt to centralize cultural event information - several local cultural sector actors in Småland had already transitioned away from using regional/national portal or database solutions. This shift was prompted by the perceived high costs and risks of obsolescence of centralized systems. As a result, institutions made significant investments in new commercial services and local solutions which they perceived as effectively meeting their needs for digital marketing and event communication. Moreover, that investment into local solutions and organizational-managed social media channels left no further resources to be allocated to a novel venture. Therefore, the participants expressed concerns regarding the hosting, maintenance, and funding of a joint platform aggregating regional cultural events. Nor did they express any interest in assuming responsibility for these aspects: as one participant simply put it, '*what's in it for me*?'

Interestingly, the focus group saw the virtue of the fragmented information landscape, expressing concern that a platform aggregated cultural events information from across the region would risk an information overload for end-users. It was doubted that the platform could appropriately filter an expansive list of events for the general public, or the public's ability to apply such filtering mechanisms. But there was another value seen in the fragmented and siloed cultural sector: the internal competition within Småland. The participants described their experience using the region's joint service for international marketing, and noted that once tourists arrived, the local actors were more likely to view one another as competitors for visitors and audiences rather than collaborators harmonious working towards a common and idealistic goal (e.g., '*cultural enrichment*'). There was therefore a pragmatic self-interest which also motivated the current, fragmented information ecosystem.

Analysis

Labor and maintenance were main concerns among the stakeholders of Project X. While the pro-innovation discourse focuses the development of always newer products, within CIS there is a growing attention to maintenance and repair (Graham & Thrift, 2007; Russel & Vinsel, 2018). This strand of research questions the innovation-centred discourse of technological development by highlighting the importance of repairing infrastructures and the work of maintaining digital

services. Project X offered to develop a free platform for circulating information about cultural events. However, it became evident that the primary costs associated with disseminating information did not lie in the development of a digital tool, but rather in its ongoing maintenance and the continuous input of information. Even if this information is free, the creation of it is cumbersome and costly. In the context of digital technologies, much of the labor required for a service to function is often obscured and black-boxed. Data needs to be cleaned, edited and errors checked (cf. Jarlbrink, 2020). Such digital labor was left out of the promises of Project X but was recognized by the cultural sector actors who were reluctant to take on any additional tasks beyond their existing efforts dedicated to local solutions and social media activities, which were themselves implemented due to the inadequacy of previous innovations.

Furthermore, the social challenge that the innovation of Project X sought to solve was not a challenge identified by important stakeholders. The idea of a joint platform for cultural events in Småland was built on an implicit comprehension within Project X of this region as an imagined community with mutual goals and interest in terms of sharing and disseminating information about culture. However, Småland encompasses a vast area, comprising three counties, two provinces (with Öland being part of Kalmar County), and 33 municipalities, each characterized by significant local variations and numerous potential cultural sector actors operating at various levels. These actors, in turn, cater to diverse audiences and pursue differing, sometimes conflicting objectives with respect to a digital platform for sharing information about cultural events, often prioritizing communities outside of Småland as their primary focus. Whereas Project X built the innovation on a presumption of collaboration, the actors were just as much competitors, or coopetitors (e.g., Gnyawali & Park, 2010).

Discussion and conclusion

Research and universities are distinctly embedded within the pro-innovation discourse. The research policy direction at both Swedish national and EU levels clearly delineates this trajectory, further reinforced by the growing dependence on external funding, which prioritizes innovation-oriented research projects, often aimed at producing and commercializing a product. In this context, research outcomes are framed in terms of innovations that are to be capitalized upon by being converted into continually new and improved services and products. Project X, as analysed in this paper, was born out of such policies. It was also funded by a funding agency that specifically endorses values deeply rooted in the pro-innovation discourse. Hence, it is not surprising that the project goals sought to fit the market- and product-oriented approach associated with such values and ideals. This resulted in attempts to develop a one-product-fits-all digital platform along with refining a best practice for information-sharing presumed to be universally applicable across Småland. Hence, the view ingrained in the pro-innovation discourse made it easy to overlook specific conditions and contexts (cf. Fougère et al., 2017).

The assumption that the digitization of objects, practices, and processes is inherently beneficial and will lead to enhanced efficiency and overall improvement shaped the direction of Project X. Yet ongoing cultural changes and current adaptations must be understood to ensure a new digital solution is actually needed and this requires looking passed a linear, deterministic trajectory dictated by solely technological advancements. The costly maintenance of novel products and services is all too often black-boxed in the '*innovationism*' (Valaskivi 2012) of today.

The market- and product-oriented approach towards innovation involves a temporal dimension that influences the focus and outcomes of research, and Project X is hardly exceptional in this regard. Depicting innovation as a competition or a race, characterized by an imagined need to maintain a high pace puts an emphasis on speed. In Project X, such incentives meant that slower processes allowing for relationship-building, reflection, and deeper insights were set aside. By critically reflecting on this process, we have sought to further on avoid reproducing the top-down innovation practice described by Varoufakis (2024), where innovations are created by a select

privileged few for consumption by the many. As researchers we must pay enough attention to relevant local conditions and key contextual aspects by critically engaging with the innovation process of the project itself. In this analysis of Project X, we have sought to reverse the top-down approach, to recognize the knowledge contributed by the various participating actors (cf. Fornstedt, 2021), and not allow instances of innovation hesitancy get overshadowed by pro-innovation bias or selectively attending only to the affirmations from more supportive stakeholders. Thereby, the project has yielded valuable insights that can be of use for understanding the actual needs of cultural sector actors with regards to digital innovation and information sharing. Nevertheless, this issue warrants further discussion and exploration.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable feedback.

About the authors

Hanna Carlsson is Associate Professor in Library and Information Science at Linnaeus University, Sweden. She received her Ph.D. from Lund University, Sweden, and her research interests are in critical studies of information literacy, digital transformation, and libraries. She can be contacted at hanna.carlsson@lnu.se

Ahmad Kamal is Senior lecturer in Library and Information Science and Digital Humanities at Linnaeus University, Sweden. He received his Ph.D. from Western University; Canada and his research interests include information practices and knowledge organization. He can be contacted at ahmad.kamal@lnu.se

Charlie Järpvall is Senior lecturer in Library and Information Science at Linnaeus University, Sweden. He received his Ph.D. from Lund University, Sweden, and his research interests include media history and information history. He can be contacted at charlie.jarpvall@lnu.se

References

Du, T., Xie, I., & Waycott, J. (2020). Marginalized communities, emerging technologies, and social innovation in the digital age: Introduction to the special issue, Information Processing & Management, 57(3), 102235. https://doi.org/10.1016/j.ipm.2020.102235.

Fornstedt, H. (2021). Innovation Resistance: Moving Beyond Dominant Framings. Diss. Uppsala: University of Uppsala, 2021. Uppsala.

Fougère M. & Harding, N. (2012). On the Limits of What Can Be Said about 'Innovation': Interplay and Contrasts Between Academic and Policy Discourses, pp 15–36 in Sveiby, K., Gripenberg, P. & Segercrantz, B. (eds.) (2012). Challenging the innovation paradigm. New York: Routledge.

Fougère, M., Segercrantz, B., & Seeck, H. (2017). A critical reading of the European Union's social innovation policy discourse: (Re)legitimizing neoliberalism. Organization, 24(6), 819-843. https://doi.org/10.1177/1350508416685171

Gnyawali, D. R., & Park, B.-J. (Robert). (2011). Co-opetition between giants: Collaboration with competitors for technological innovation. Research Policy, 40(5), 650–663. https://doi.org/10.1016/j.respol.2011.01.009

Godin B. & Vinck D. (2017). Introduction: innovation – from the forbidden to a cliché, pp 1–14 in Godin, B. & Vinck, D. (eds.) (2017). Critical studies of innovation: alternative approaches to the pro-innovation bias. Cheltenham, UK: Edward Elgar Publishing.

Graham, S., & Thrift, N. (2007). Out of Order: Understanding Repair and Maintenance. Theory, Culture & Society, 24(3), 1-25. https://doi.org/10.1177/0263276407075954

Hallonsten, O. (2023). Empty Innovation: Causes and Consequences of Society's Obsession with Entrepreneurship and Growth (1st ed.). Springer Nature. https://doi.org/10.1007/978-3-031-31479-7

Heidenreich, S., & Kraemer, T. (2016). Innovations-Doomed to Fail? Investigating Strategies to Overcome Passive Innovation Resistance. The Journal of Product Innovation Management, 33(3), 277–297. https://doi.org/10.1111/jpim.12273

Jarlbrink, J. (2020). All the work that makes it work: Digital methods and manual labour. In Digital historie : Emergent approaches within the new digital history (1st ed., pp. 113–126). https://doi.org/10.33134/HUP-5-7

Pfotenhauer, S., & Jasanoff, S. (2017). Panacea or diagnosis? Imaginaries of innovation and the 'MIT model' in three political cultures. Social Studies of Science, 47(6), 783–810. https://doi.org/10.1177/0306312717706110

Russell, A.L., & Vinsel, L. (2018). After Innovation, Turn to Maintenance. Technology and Culture 59(1), 1-25. https://dx.doi.org/10.1353/tech.2018.0004.

Stucki, T. (2019). What hampers green product innovation: the effect of experience. Industry and Innovation, 26(10), 1242–1270. https://doi.org/10.1080/13662716.2019.1611417

Thomä, J. (2017). DUI mode learning and barriers to innovation—A case from Germany. Research Policy, 46(7), 1327–1339. https://doi.org/10.1016/j.respol.2017.06.004

Valaskivì, K. (2012). Dimensions of innovationism. In Nynäs, P., Lassander, M. & Utriainen, T. (eds). Post-secular society. New Brunswick, New Jersey: Transaction Publishers (pp. 129–156).

Varoufakis, Y. (2024). Technofeudalism: What killed capitalism. Melville House.

Wormbs, Nina (2010). Det digitala imperativet. Efter The Pirate Bay, pp 140-150 in Andersson, J. & Snickars, P. (eds.) (2010). Efter the Pirate Bay. Stockholm: Kungliga biblioteket.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

How does digital humanities research talk about AI? A bibliometric analysis

Wenyi Shang, Rongqian Ma, and Heather Moulaison-Sandy

DOI: https://doi.org/10.47989/ir30iConf47242

Abstract

Introduction. Despite scholarly work that looked into artificial intelligence (AI)'s impact on digital humanities (DH), there has been little comprehensive investigation into how AI has been reflected and adopted in DH scholarship. This paper addresses this gap by asking: What patterns can be observed in AI integration in the DH scholarly discourse?

Method. A bibliometric analysis was conducted on 2,488 abstracts of articles published in three influential DH journals: *digital scholarship in the humanities*, *digital humanities quarterly*, and *journal of cultural analytics*.

Analysis. Focusing on abstracts containing the term 'AI' or '*artificial intelligence*,' this study conducted four analyses: (1) temporal analysis of AI mentions, (2) collocation analysis, (3) word vector analysis, and (4) topic modeling.

Results. AI has been discussed in DH scholarship well before its recent rise in prominence. The recurring themes in these discussions encompass both technological and human-centred aspects.

Conclusion. AI has been integral to DH, and the discourse around AI in DH reflects the field's dual focus on technology and the humanities. The bibliometric analysis presented in this paper illustrates how information science can inform and guide methodological reflections in DH, offering new insights into the future development of the field.

Introduction

Advances in computing technology have brought changes to multiple aspects of humanities scholarship and research culture, including data practices, methodologies, knowledge production practices, as well as scholarly communication processes (Berry, 2022). In particular, the recent surge in generative artificial intelligence (AI) has garnered attention for its ability to automate creative processes, posing new challenges and opportunities for scholars in digital humanities (DH) while creating a valuable moment to revisit the long history of AI in DH (Gefen et al., 2021). Despite scholarly works that looked into AI's impact on DH both conceptually and empirically, there has been little comprehensive investigation into how AI has been reflected and adopted in DH scholarship. This type of examination of AI adoption in DH, however, can be of great importance, offering insights into, and guidance on, the future integration of AI in DH.

Our paper aims to address this gap by asking: What patterns can be observed in AI integration in DH scholarly discourse? Using a bibliometric approach, we conduct an exploratory analysis of AI terms ('AI' and '*artificial intelligence*') in article abstracts published across three influential DH journals: *digital scholarship in the humanities* (*DSH*, formerly known as *literary and linguistic computing*), digital *humanities quarterly* (*DHQ*), and *journal of cultural analytics* (*JCA*). More specifically, our analysis answers the following questions:

(1) What is the temporal trend of AI mentions represented in the abstracts?

(2) How do AI terms collocate with other concepts in these abstracts?

(3) What are the semantic underpinnings of AI when discussed in these abstracts?

(4) What are the major topics or domains within DH that are at the forefront of AI mentions, as seen in the abstracts?

By investigating AI mentions in DH scholarship, our work aims to empirically illustrate an evolving picture of AI integration in DH, bridging the historical and contemporary practices.

Literature review

AI integration in DH

The integration of AI technology and AI-based tools (e.g., machine learning (ML), natural language processing (NLP)) has accelerated the '*computational turn*' of the field of DH, expanding its capacity to handle more complex datasets, uncover hidden patterns, and automate processes that enhanced both the scale and depth of the scholarly inquiry (Berry, 2022; Gefen et al., 2021). In text analysis, for instance, AI-powered tools, such as NLP, have facilitated the examination of large text corpora, enabling the practice of '*distant reading*' (Moretti, 2013). Furthermore, a growing focus on data and algorithms—particularly on how to conceptualize and work with data, and how to understand as well as address the opaque nature of algorithmic processes and their underlying data transformations—has brought important questions to the forefront of DH research (Bode, 2019; D'Ignazio & Klein, 2020; Drucker, 2011; Lavin, 2021; Schmidt, 2016). In addition to enabling methodological transformation, AI-driven automation has reshaped how labor is conceptualized and executed in DH, introducing new opportunities for interdisciplinary collaborations and research innovation (Smithies, 2017). Building on AI's longstanding presence in DH, the recent rise of generative AI technology presents both new possibilities and challenges for the field. Researchers have begun exploring how this technology can be integrated into DH research and pedagogical practices in ways that are effective, critical, and ethical (Chun & Elkins, 2023; Dedema & Ma, 2024; Guo, 2024; Liu et al., 2023).

Despite the recent progress in exploring AI applications within DH and the acknowledgement of the importance of data and related ethical implications in DH research (Gonzalez & Rodrigues,

2022), a comprehensive, empirical investigation of how AI has been reflected in DH scholarship remains limited. This is especially true when considering the historical integration of AI technologies into DH publications and research. With the recent surge in generative AI technologies, a historical perspective on AI's use in DH is increasingly valuable for identifying patterns, gaps, and emerging challenges within the field. Our paper responds to this gap by examining AI use in DH scholarship, tracing its evolution from the 1980s to the present.

Bibliometric analysis of DH

Bibliometric analysis offers valuable methods for illustrating broad trends and usage patterns in DH scholarship through empirical data. Previous studies have highlighted the strengths of bibliometric approaches—such as citation analysis, network analysis, and topic modeling—in examining DH as an interdisciplinary field and research community. For instance, Luhmann and Burghardt (2021) analysed articles from DH journals alongside those from 15 other academic disciplines, addressing the conceptual identity of DH—one of the field's enduring debates—and underscored how DH connects with neighbouring fields like computational linguistics and information science. Similarly, Ma and Li (2021) introduced the '*battleground*' metaphor, emphasizing the tensions and collaborations between humanities and STEM research practices and conventions. Earlier works demonstrated that bibliometric analysis also sheds light on the early history and emergence of DH as a field (Leydesdorff & Akdag Salah, 2010; Sula & Hill, 2019), as well as its gradual evolution into a '*Big Tent*' discipline (Weingart & Eichmann-Kalwara, 2017). Furthermore, citation and network analysis using DH journal article data has provided insights into global DH research collaboration patterns and community structures (Gao et al., 2017, 2018; Li et al., 2024; Nyhan & Duke-Williams, 2014; Tang et al., 2017).

In this paper, we leverage the unique methodological advantages of bibliometric analysis and apply it to examine a publication corpus of three influential DH journals, highlighting the broad patterns of AI adoption in DH scholarship.

Methods

Data collection

We collected metadata from all articles published in three influential, '*exclusively DH*' journals: *DSH*, *DHQ*, and *JCA* (Spinaci et al., 2020). Published since 1986, *DSH* allows us to consider nearly 40 years of DH history. Metadata for *DSH* articles were parsed from HTML files collected by one of the authors using their university library's subscription and access to the journal; the *DHQ* metadata were extracted from XML files available in the journal's GitHub repository (https://github.com/Digital-Humanities-Quarterly/dhq-journal/tree/main/articles). For *JCA*, the metadata were obtained by web scraping the HTML source code of the journal's online open-access articles.

In this study, we analysed abstracts rather than full-text articles because abstracts provide higher information density. We expect that articles focused on AI will include relevant terms in their abstracts. Therefore, we excluded articles without abstracts or publication years, as well as those with metadata information indicating that the articles are not in English. This resulted in a corpus of 1,587 abstracts from *DSH* (69.3% of the total 2,290 articles published between 1986 and 2023), 720 abstracts from *DHQ* (95.2% of the total 756 articles published between 2007 and 2024), and 181 abstracts from *JCA* (99.5% of the total 182 articles published between 2016 and 2024). After data collection, four data analysis experiments were conducted to examine the discourse surrounding AI in these abstracts: temporal analysis of AI mentions, collocation analysis, word vector analysis, and topic modeling.

Temporal analysis of AI mentions

For the temporal analysis of AI mentions, we measured the presence of AI terms in each abstract and tracked their variation over time. We defined '*AI terms*' as 'AI' and '*artificial intelligence*' only, rather than creating an extended thesaurus, to ensure that all terms discussed in this study specifically refer to AI. Abstracts were identified as containing AI mentions if they included either the word '*ai*' or the 2-gram '*artificial intelligence*' (where '*artificial*' is immediately followed by '*intelligence*' within the same sentence). We then calculated the proportion of articles with AI mentions in their abstracts published each year.

Collocation analysis

Collocation analysis was conducted to identify the meaningful terms that co-occur with AI terms in each abstract. We first generated a word count table using the CountVectorizer from the Python module Scikit-learn (Pedregosa et al., 2011). To reduce morphological variants of words to a common stem, we applied the SnowballStemmer for English from the NLTK library (https://www.nltk.org/api/nltk.stem.SnowballStemmer.html). Next, we calculated the number of abstracts in which each stemmed word co-occurs with AI terms. Following the ideas of Algee-Hewitt et al. (2020) and Gilkison and Kurzynski (2024), we employed Fisher's exact test (Fisher, 1922) to identify statistically significant collocates of AI terms, excluding common words like '*the*' and '*a*'. Additionally, referencing Gilkison and Kurzynski (2024), we calculated the ratio of the actual number of abstracts where each stemmed word co-occurs with AI terms to the expected number of such co-occurrences (r-exp), which is defined as:

$$r - exp_i = \frac{\#\ of\ Abstracts\ Containing\ the\ Stemmed\ Word\ i\ and\ AI\ Terms}{\#\ of\ Abstracts\ Containing\ the\ Stemmed\ Word\ i \cdot \frac{\#\ of\ Abstracts\ Containing\ AI\ Terms}{Total\ \#\ of\ Abstracts}}$$

Word vector analysis

In addition to words that co-occur with AI terms, we also identified words with similar semantic meanings with AI terms through word vector analysis, which quantifies the semantic similarity between terms based on their contextual usage. Using the Python library Gensim (Řehůřek & Sojka, 2010), we pre-processed the corpus with simple_process and converted each stemmed word into a vector with the Word2Vec model (Mikolov et al., 2013). For better performance with our relatively small dataset, we adjusted the '*min_count*' parameter to 1 to ensure no word was excluded, and switched the '*sg*' parameter from 0 to 1 to use the skip-gram algorithm instead of the CBOW algorithm. We then calculated the cosine similarity between each stemmed word's vector and the vector of the word '*ai*', ranking their semantic similarity based on these cosine similarity scores.

Topic modeling

Topic modeling is a technique used to '*discover hidden thematic structure in large collections of texts*', and is widely used in DH (Blei, 2012). For this study, we applied topic modeling to examine the themes in abstracts and their relationship with AI terms. We used the MALLET (MAchine Learning for LanguagE Toolkit) for this purpose, which offers efficient sampling-based implementations of latent Dirichlet allocation (LDA), Pachinko allocation, and hierarchical LDA in its topic modeling toolkits (https://github.com/mimno/Mallet) and is commonly used by digital humanists (e.g., Graham et al., 2012). We imported our corpus into MALLET and generated 30 topics. When importing data into MALLET, we specified the '*--token-regex*' as '[\p{L}] +,' since the default regular expression removes words with 1 or 2 characters, including "ai." We also combined the default English stop word list with the French stop word list from https://github.com/stopwords-iso/stopwords-fr (excluding the word '*ai*'). The French stop word list was necessary because some abstracts contain French words without metadata indicating their language, making it impossible to filter them out during data collection. In determining the number of topics, we tested various topic numbers in increments of 10 (from 10 to 50) and observed that

the log-likelihood initially rose, then declined, reaching its peak at 30. We then compared the average weight of each topic in abstracts with AI terms to those without. Finally, the three co-authors independently labelled each topic based on the 20 keywords for each topic, and then synthesized their results to determine the final label for each topic.

Results

Temporal analysis of AI mentions

Out of a total of 2,488 abstracts, 28 (1.1%) contain AI terms. The temporal variation in the proportion of articles with AI mentions in their abstracts published each year is illustrated in Figure 1.

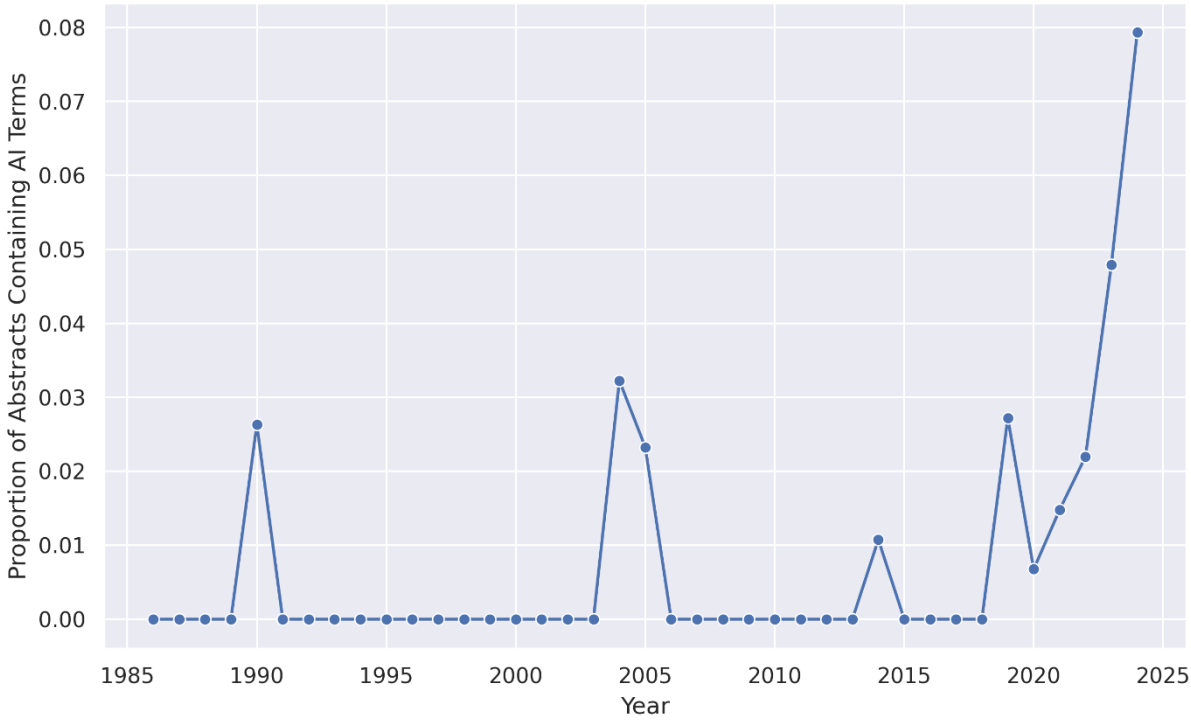


Figure 1. Temporal variation of AI mentions in the abstracts of *DSH*, *DHQ*, and *JCA*

Overall, the leading DH journals show a sharp increase in AI mentions in their abstracts after 2019, with a particularly notable surge beginning in 2023. This trend aligns with the recent development of AI technology, especially generative AI, which has drawn significantly increased attention from DH scholars, potentially leading to more extensive discussions in these journals. Additionally, it is noteworthy that four articles mentioning AI terms were published before this recent rise. With the earliest dating back to 1990, they applied AI to tasks including text interpretation (De Vuyst, 1990; Terras, 2005), text encoding (Terras & Robertson, 2004), and optical character recognition (OCR) (Heckmann et al., 2014). This finding supports Berry's (2022) observation of digital humanists' extended history of using automation (pp. 446–447), which predates the recent surge of AI in public discourse, and further highlights that the term 'AI' was occasionally explicitly used in these early practices.

Collocation analysis

There are 16 stemmed words identified as statistically significant collocates of AI terms ($p < 0.001$ in Fisher's exact test, excluding '*ai*', '*artifici*', and '*intellig*'). These collocates are presented in Table 1, ranked in ascending order of p-values (i.e., more statistically significant collocates appear higher in the list).

Rank	Word	r-exp	p-value
1	ethic	15.55	1.49E-07
2	system	3.29	1.45E-05
3	learn	4.21	5.24E-05
4	big	11.11	6.23E-05
5	machin	4.62	6.74E-05
6	dungeon	88.86	1.22E-04
7	big-fiv	88.86	1.22E-04
8	loop	88.86	1.22E-04
9	openai	88.86	1.22E-04
10	gpt-2	88.86	1.22E-04
11	digitis	22.21	2.63E-04
12	capabl	8.08	2.93E-04
13	recognit	5.92	3.76E-04
14	offlin	44.43	7.23E-04
15	historian	6.53	7.93E-04
16	ecolog	14.81	9.31E-04

Table 1. Statistically significant collocates of AI terms

The 16 statistically significant collocates of AI terms span both technical and humanities-oriented concepts. Technical terms like '*system*', '*machin*', '*openai*', and '*gpt-2*' highlight the technical focus in discussions of AI within DH scholarship. Simultaneously, humanities-oriented terms such as '*historian*' indicate the application of AI in specific humanities disciplines and underscore the prominence of humanistic perspectives in DH work. Additionally, the presence of terms such as '*ethic*' further highlights the field's emphasis on ethical issues. This observation is also echoed in digital humanities scholarship (e.g., Gonzalez & Rodrigues, 2022), especially in the context of teaching with emerging AI technologies (e.g., Chun & Elkins, 2023; Berry, 2022).

Word vector analysis

Among the 12,418 distinct stemmed words (excluding '*ai*' itself), it is unsurprising that many of the terms with the highest semantic similarity to 'ai' are technical. For example, '*intellig*' (5th), '*portal*' (6th), and '*microcomput*' (11th) were among the terms ranked the highest in semantic similarity to '*ai.*' This potentially demonstrates an alignment between DH and information science in perceiving these technical elements as integral to the study of AI.

Topic modeling

As shown in Figure 2, the 28 abstracts containing AI terms have substantially higher average weights (at least twice as high) in topics 10 (Ancient Texts), 17 (Digital Cultural Heritage), 23 (Narratology), 24 (Machine Learning), and 25 (Digital Media) compared to the 2,460 abstracts without AI terms.

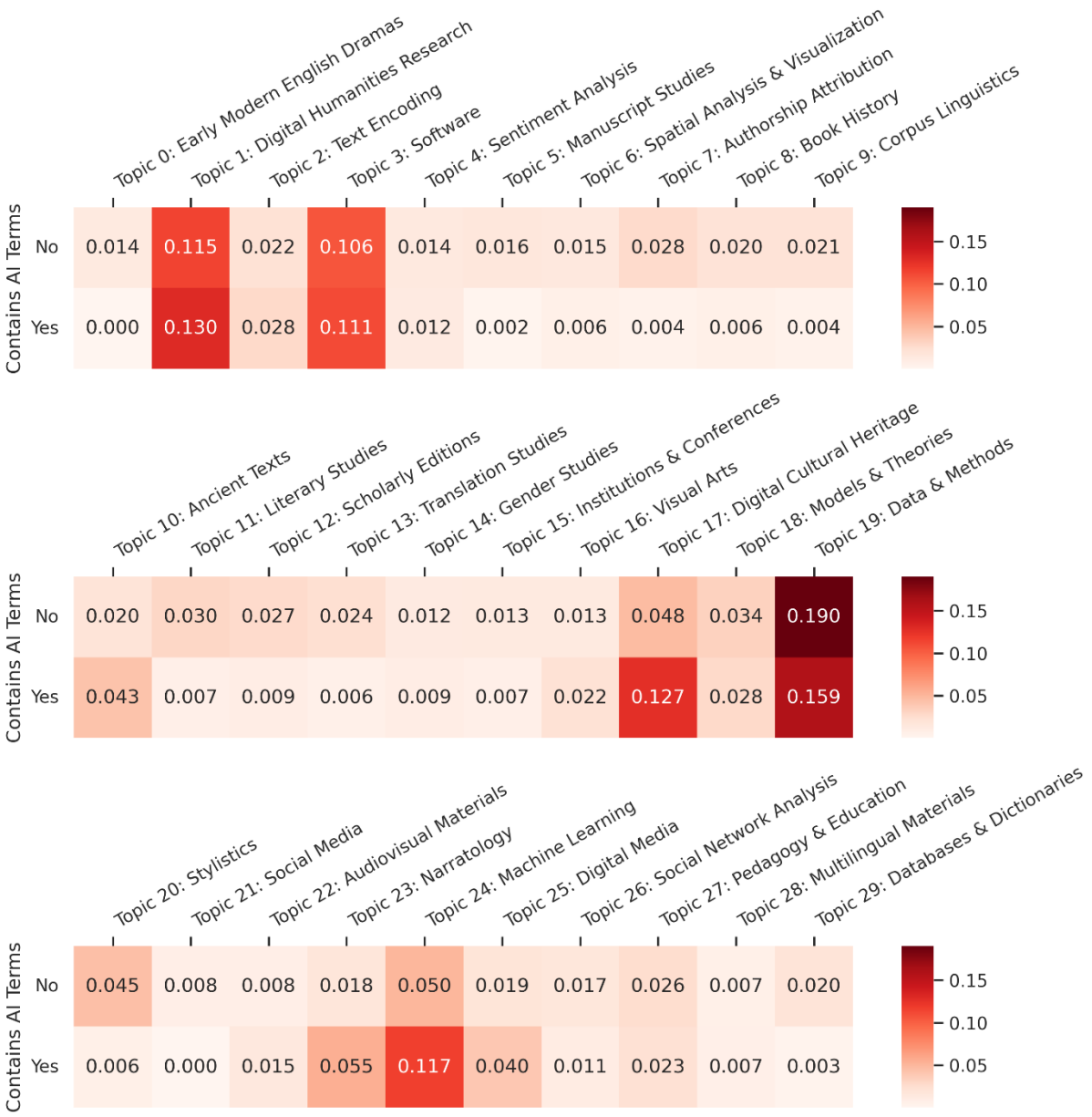


Figure 2. Average topic weights for abstracts with and without AI terms

Topics associated with AI in the topic modeling process aligned with themes previously revealed in the analysis (i.e., technical and humanities-oriented). Technical aspects are strongly demonstrated in Topic 24, which is comprised of keywords related to ML methodology (e.g., '*machine*', '*learning*', '*accuracy*', '*classification*', '*features*'). Articles with abstracts where Topic 24 is dominant (weight > 0.4) reflect a particular interest in technical aspects, including the application of AI and ML to tasks like OCR (Heckmann et al., 2014), text classification (Zheng & Jin, 2019), and information extraction (Sánchez-Salido et al., 2023).

AI is also aligned with humanities-oriented themes. Topics 10, 17, 23, and 25 each represent an intersection between DH and a specific humanities field. Topic 10 highlights keywords related to ancient texts and language systems (e.g., '*Greek*', '*ancient*', '*language*', '*system*', '*linguistic*'); Keywords of Topic 17 centres on cultural heritage (e.g., '*cultural*', '*heritage*', '*historical*', '*archive*', '*collection*'); those of Topic 23 focuses on narratology (e.g., '*narrative*',' *character*', '*story*', '*novel*', '*plot*'); and those of Topic 25 concerns media studies (e.g., '*media*', '*video*', '*game*', '*art*', '*virtual*'). This

suggests that certain humanities disciplines—linguistics, heritage science, media studies, and a subfield of literature, narratology—may be at the forefront of incorporating AI advancements, reflecting their active engagement in AI-driven research.

Discussion and conclusion

In summary, the results of this empirical study reveal that AI has been discussed in DH scholarship well before its recent rise in prominence. These discussions encompass a wide range of perspectives, from technical approaches to those aligned with specific humanities disciplines, as well as efforts by information professionals to facilitate access to collections through ethical engagement and professional practice. The recurring themes of both technological and human-centred aspects present no surprise, given the dual focus of the DH field on technology and the humanities.

However, it is surprising that only 1.1% of abstracts in the corpus explicitly mention 'AI' or '*artificial intelligence*'. Even when considering only publications from the 2020s—a period when AI has become an extremely prominent topic in public discourse—this proportion increases only slightly to 2.7%. This may lie precisely in the close association between AI and the computing technologies that enable DH work—DH is so inherently linked to AI that explicit mentions of it might seem unnecessary. For authors, the need to explicitly identify AI technologies might not be as strongly felt in this fundamentally technology-based field.

The low proportion of abstracts mentioning AI terms can also be explained by the possibility that, in some cases, AI is considered as synonymous with the computational aspects integral to DH. In future work, we plan to consider additional AI-related terms (e.g., ML, NLP, large language models) to provide a comprehensive view of the use of AI technologies. Future research could also use the full texts of articles and a wider range of DH journals. This approach would enable a more comprehensive and nuanced interpretation of AI integration in DH, complementing the insights gained from abstract analysis.

In conclusion, this paper represents preliminary work to address the gap in empirical research on AI adoption in DH scholarship. Our findings demonstrate that algorithmic processes have significantly impacted the landscape of DH research, with practitioners actively leveraging recent advancements in AI technology. This underscores the importance of reflecting on AI integration in DH. Our bibliometric analysis illustrates how information science can inform and guide such methodological reflections in DH, offering new insights into the future development of the field.

About the authors

Wenyi Shang is Assistant Professor in the iSchool, University of Missouri, USA. He received his Ph.D. from the University of Illinois Urbana-Champaign, and his research interests are in the digital humanities. He can be contacted at wenyishang@missouri.edu.

Rongqian Ma is Assistant Professor of Information and Library Science in the Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, USA. She received her Ph.D. from the University of Pittsburgh, and her research interests focus on computational literary studies, digital archives, and scholarly communication. She can be contacted at rm56@iu.edu.

Heather Moulaison-Sandy is Associate Professor in the iSchool, University of Missouri, USA. She received her Ph.D. from Rutgers University, and her research interests focus on the organization of information, scholarly communication, and technology. She can be contacted at moulaisonhe@missouri.edu.

References

Algee-Hewitt, M., Porter, J. D., & Walser, H. (2020). Representing race and ethnicity in American fiction, 1789-1920. Journal of Cultural Analytics, 5(2). https://doi.org/10.22148/001c.18509

Berry, D. M. (2023). AI, ethics, and digital humanities. In J. O'Sullivan (Ed.), The Bloomsbury handbook to the digital humanities (pp. 445–457). Bloomsbury Academic.

Blei, D. M. (2012). Topic modeling and digital humanities. Journal of Digital Humanities, 2(1). https://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei

Bode, K. (2019). A world of fiction: Digital collections and the future of literary history. University of Michigan Press.

Chun, J., & Elkins, K. (2023). eXplainable AI with GPT4 for story analysis and generation: a novel framework for diachronic sentiment analysis. International Journal of Digital Humanities, 5(2), 507–532. https://doi.org/10.1007/s42803-023-00069-8

De Vuyst, J. (1990). Knowledge representation for text interpretation. Literary and Linguistic Computing, 5(4), 296–302. https://doi.org/10.1093/llc/5.4.296

Dedema, M., & Ma, R. (2024). The Collective Use and Evaluation of Generative AI Tools in Digital Humanities Research: Survey-Based Results (arXiv:2404.12458). arXiv. https://doi.org/10.48550/arXiv.2404.12458

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. The MIT Press.

Drucker, J. (2011). Humanities approaches to graphical display. Digital Humanities Quarterly, 5(1). https://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Fisher, R. A. (1922). On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society, 85(1), 87–94. https://doi.org/10.2307/2340521

Gao, J., Duke-Williams, O., & Mahony, S. (2017). The intellectual structure of digital humanities: an author co-citation analysis. The Book of Abstracts of Digital Humanities Conference 2018, 450–452.

Gao, J., Nyhan, J., Duke-Williams, O., & Mahony, S. (2018). Visualising the digital humanities community: a comparison study between citation network and social network. The Book of Abstracts of Digital Humanities Conference 2018, 201–204.

Gefen, A., Saint-Raymond, L., & Venturini, T. (2021). AI for digital humanities and computational social sciences. In B. Braunschweig & M. Ghallab (Eds.), Reflections on artificial intelligence for humanity (pp. 191–202). Springer. https://doi.org/10.1007/978-3-030-69128-8_12

Gilkison, A., & Kurzynski, M. (2024). Vectors of violence: legitimation and distribution of state power in the People's Liberation Army Daily (Jiefangjun Bao), 1956-1989. Journal of Cultural Analytics, 9(1). https://doi.org/10.22148/001c.115481

Gonzalez, M. E., & Rodrigues, M. V. (2022). Digital humanities: ethical implications and interdisciplinary challenges. Humanities Bulletin, 5(1), Article 1.

Graham, S., Weingart, S., & Milligan, I. (2012, September 2). Getting started with topic modeling and MALLET. Programming Historian. https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Guo, Q. (2024). Prompting Change: ChatGPT's impact on digital humanities pedagogy: a case study in art history. International Journal of Humanities and Arts Computing, 18(1), 58–78. https://doi.org/10.3366/ijhac.2024.0321

Heckmann, D., Frank, A., Arnold, M., Gietz, P., & Roth, C. (2014). Citation segmentation from sparse & noisy data: a joint inference approach with Markov Logic Networks. Digital Scholarship in the Humanities, 31(2), 333–356. https://doi.org/10.1093/llc/fqu061

Lavin, M. (2021). Why digital humanists should emphasize situated data over capta. Digital Humanities Quarterly, 15(2). https://digitalhumanities.org/dhq/vol/15/2/000556/000556.html

Leydesdorff, L., & Salah, A. A. A. (2010). Maps on the basis of the Arts & Humanities Citation Index: The Journals Leonardo and Art Journal versus 'Digital Humanities' as a topic. Journal of the American Society for Information Science & Technology, 61(4), 787–801. https://doi.org/10.1002/asi.21303

Li, K., Ma, R., & Fang, Z. (2024). Explicit or implicit digital humanities? An examination of search strategies to retrieve digital humanities publications from large-scale scholarly databases. Quantitative Science Studies. Advance online publication. https://doi.org/10.1162/qss_a_00321

Liu, C., Wang, D., Zhao, Z., Hu, D., Wu, M., Zhang, H., Lin, L., Liu, J., Shen, S., Li, B., & Zhao, L. (2024). SikuGPT: a generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. ACM Journal on Computing and Cultural Heritage. Advance online publication. https://doi.org/10.1145/3676969

Luhmann, J., & Burghardt, M. (2021). Digital humanities—a discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. Journal of the Association for Information Science and Technology, 72(12), 1475-1498. https://doi.org/10.1002/asi.24533

Ma, R., & Li, K. (2021). Digital humanities as a cross-disciplinary battleground: an examination of inscriptions in journal publications. Journal of the Association for Information Science and Technology, 73(2). https://doi.org/10.1002/asi.24534

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings. http://arxiv.org/abs/1301.3781

Moretti, F. (2013). Distant reading. Verso.

Nyhan, J., & Duke-Williams, O. (2014). Joint and multi-authored publication patterns in the digital humanities. Literary and Linguistic Computing, 29(3), 387–399. https://doi.org/10.1093/llc/fqu018

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-Learn: machine learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50.

Sánchez-Salido, E., Menta, A., & García-Serrano, A. (2023). Seeking information in Spanish historical newspapers: the case of Diario de Madrid (18th and 19th Centuries). Digital Humanities Quarterly, 17(4). https://www.digitalhumanities.org/dhq/vol/17/4/000735/000735.html

Schmidt, B. (2016). Do digital humanists need to understand algorithms? In M. K. Gold & L. F. Klein (Eds.), Debates in the digital humanities 2016. University of Minnesota Press.

Smithies, J. (2017). The digital humanities and the digital modern. Palgrave Macmillan.

Spinaci, G., Colavizza, G., & Peroni, S. (2020). Preliminary Results on Mapping Digital Humanities Research. In C. Marras, M. Passarotti, G. Franzini, & E. Litta (Eds.), La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica: Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD): 15-17 gennaio 2020, Milano, Università Cattolica del Sacro Cuore, 246–252.

Sula, C. A., & Hill, H. V. (2019). The early history of digital humanities: an analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004). Digital Scholarship in the Humanities, 34(Supplement_1), i190–i206. https://doi.org/10.1093/llc/fqz072

Tang, M.-C., Cheng, Y. J., & Chen, K. H. (2017). A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses. Scientometrics, 113(2), 985–1008. https://doi.org/10.1007/s11192-017-2496-6

Terras, M. (2005). Reading the readers: modelling complex humanities processes to build cognitive systems. Literary and Linguistic Computing, 20(1), 41–59. https://doi.org/10.1093/llc/fqh042

Terras, M., & Robertson, P. (2004). Downs and acrosses: textual markup on a stroke level. Literary and Linguistic Computing, 19(3), 397–414. https://doi.org/10.1093/llc/19.3.397

Weingart, S. B., & Eichmann-Kalwara, N. (2017). What's under the big tent?: A study of ADHO conference abstracts. Digital Studies, 7(1), 6. https://doi.org/10.16995/dscn.284

Zheng, W., & Jin, M. (2019). Comparing multiple categories of feature selection methods for text classification. Digital Scholarship in the Humanities, 35(1), 208–224. https://doi.org/10.1093/llc/fqz003

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

The dual impact of digital humanities: evaluating the role of DH approaches in historical and cultural resource development

Lu Sijia, Long Jiaqing, and Xu Yongjun

DOI: https://doi.org/10.47989/ir30iConf47236

Abstract

Purpose. Digital humanities (DH) approaches have irreversibly impacted historical and cultural resource development. This study critically examines the dual impact of DH approaches to highlight their contributions and limitations in resource preservation and enhancement.

Method. Based on the input-process-output (IPO) model in system theory, we proposed the source-transformation-result (STR) framework to guide our content analysis of 34 selected projects awarded and nominated for the digital humanities awards (2019-2023).

Analysis. Our content analysis reveals a paradigm shift in three critical areas: (1) source: transition from isolated individual work to integrated corpora from multiple works; (2) transformation: evolve from manual interpretative close reading to computational distant reading; (3) result: advance from primary forms to interactive visual representations.

Findings. The findings indicate that DH approaches facilitate the regeneration of new knowledge, provide structured insights into extensive datasets, and enhance user engagement. Still, DH approaches also pose risks, such as the erasure of unique details and contextual richness, reduced interpretative depth, and the oversimplification of complex data.

Value. This study underscores the necessity for a balanced approach to DH methodologies to optimize their benefits while mitigating inherent drawbacks.

Introduction

Historical and cultural resources refer to a wide range of tangible and intangible things of human creation and use that occurred in the past and continue to play an essential role in modern society (Huang & Lu, 2015). Developing these resources confers substantial value, including enhancing heritage preservation and reinforcing public cultural identity. Digital humanities (DH) approaches are a powerful suite of tools and methodologies for exploring new ways of being in a digital age (Berry & Fagerjord, 2017), which has also irreversibly impacted historical and cultural resource development. They encompass digital technologies and techniques—such as digitization, data management, machine learning, data analysis, visualization, and VR/AR—and the methodological frameworks that guide their application in historical and cultural resource development. However, the integration of DH approaches is not without its challenges. Discussions on the pros and cons of DH approaches for developing historical and cultural resources still need to continue. Current studies in DH have already included some critical considerations, such as maintaining linguistic diversity in digital contexts (Karrouche, 2019) and addressing data discrimination in technology applications (Hendery et al., 2019). James E. Dobson (2019) also emphasized the need to correct the unrestrained techno-utopian impulses of DH, urging a more reflective and balanced approach.

Yet, there remains a gap in discussions on the pros and cons of using DH for developing historical and cultural resources. This study aims to fill that gap by navigating the complex landscape of DH applications in historical and cultural resource development. By doing so, we seek to ensure that DH approaches contribute positively to preserving and enriching our shared heritage, thereby fostering a deeper understanding of the past while informing our present and future. The results of this study can advance the discourse on the ethics and effectiveness of DH practices and offer practical guidelines to help maintain cultural diversity and reduce biases when utilizing DH approaches.

Research design

Data collection

As a flagship event in the DH community, the digital humanities awards (DHA, http://dhawards.org/) features a wide range of DH projects with broad scope and diverse cases. Since its establishment in 2012, nearly 40 countries have had DH projects nominated for the DHA awards, demonstrating its global engagement. This makes DHA projects a valuable dataset for researching international practices in digital humanities. Resources in any language or format, including those focused on historical and cultural resource development, are eligible for nomination. The awards span various categories, such as best DH tool or suite of tools, best DH data visualisation, best DH blog post or series of posts, and best DH dataset. Notably, recipients of the best DH data visualization award are comprehensive DH projects implemented in real-world contexts, making them more suitable as cases for this research. Out of the 77 projects acknowledged in this category over the past five years (2019-2023), we have selected 34 that specifically focus on developing historical and cultural resources for content analysis.

The selection criteria were: (1) involvement in historical and cultural resource development; (2) ongoing project operation; (3) accessible website and publicly available project-specific information. The selected projects are highly representative in three aspects. Source-wise, they span 16 countries, including Australia, China, the United States, and Italy, and so on. Form-wise, they encompass both tangible and intangible historical and cultural resources. Content-wise, they cover a wide range of topics, such as historical sites and regional historical culture, family historical culture, folk tales, poetry, literature, religion, and historical artifacts (including antiques, maps, manuscripts, old newspapers, handicrafts, among others).

These projects provide a comprehensive overview of how DH approaches are applied to historical and cultural resource development on an international scale. They showcase a diverse range of DH

approaches across key domains, such as project management, data processing, and user interaction. Through a detailed content analysis of these initiatives, we can more effectively evaluate the role and impact of DH methodologies in advancing the development of historical and cultural resources.

We acknowledge that our selected cases consist exclusively of award-winning and nominated projects from the DHA, which are determined by public voting. This may introduce a selection bias, as the awards may not fully reflect project quality, but they do promote greater diversity.

Content analysis

To evaluate the role of DH approaches in historical and cultural resource development, it is essential to address a general methodological issue in digital humanities: How do we transition from the digital to meaning? (Rodríguez Díaz, Karines, Haber Guerra, Yamile, & Gómez Masjuán, Miguel Ernesto, 2021). In this study, we employ the input-process-output (IPO) model, a key framework in systems theory, to structure the evaluation of this transition. The IPO model is defined as '*inputs that lead to processes that in turn lead to outcomes*' (Ilgen, Hollenbeck, Johnson, & Jundt, 2005).

Based on this model and tailored to the context of historical and cultural resource development, we propose the source-transformation-result (STR) framework to guide our content analysis (see Fig. 1). In this framework, the source (input) refers to the raw material for development, which we analyze primarily in terms of their content, form, and quantity. The transformation (process) involves methods adopted to convert these materials into meaningful outputs, and we mainly focus on the digital technologies and techniques used in this process. The result (output) represents the outcomes generated, such as visualizations, new knowledge, or interactive experiences. Unlike traditional IPO models, the STR framework is cyclical. The knowledge generated (e.g., semantic networks or interactive timeline) can feedback into the source phase, influencing future outputs.

The STR framework enables a comprehensive evaluation of how DH approaches impact each stage of historical and cultural resource development.

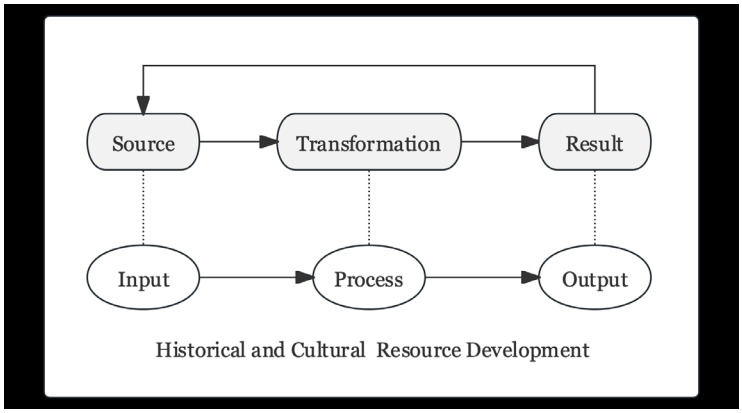


Figure 1. STR analytical framework

Findings

Through a content analysis guided by the STR framework (see Table. 1), we identified shifts in historical and cultural resource development across three dimensions attributable to the DH approaches.

(1) Source dimension: transition from isolated individual work to integrated corpora composed of data from multiple sources, which are more affluent and multi-layered. For instance, the Dong Qichang digital humanities project (P4) integrates high-definition images, work texts, and treatises

into a comprehensive digital collection. Similarly, the Gaoqian Digital Memory Website (P7) compiles diverse data types like oral interview data and architectural surveys.

(2) Transformation dimension: evolving from manual interpretative close reading to computational distant reading. Franco Moretti, who introduced distant reading, described it as a method to '*learn how not to read texts*' (Moretti, 2013), proposing employing advanced tools to manage and analyze large datasets. Projects like the Sur la piste des œuvres antiques (P2) use social network analysis and historical geographic information systems to explore historical transactions and social interactions in the French art market.

(3) Result dimension: advancing from primary forms to interactive visual representations, generating new knowledge. In this study, new knowledge is defined relatively, referring to insights and understandings that emerge from analysing and synthesizing the existing knowledge (Source). Projects such as the Moseley Homestead Digital Heritage Preservation Project (P16) employ 360 photography and high-resolution panoramic imaging for immersive exploration of historical sites, while the Surprise Machines for Harvard Art Museums (P26) leverage VR technology to create dynamic, interactive museum experiences. These advancements enhance user engagement and provide novel interpretations of the original historical and cultural resources.

No._ Project name	Source	Transformation	Result
P1 Witches – Mapping the Scottish Survey of Witchcraft Database (2019)	Different locations recorded within the Survey of Scottish Witchcraft Database	Social Network Analysis (SNA); Historical Geographic Informatisation (HGI); Chronological Analysis; Visualization	It can show a variety of different visualisations of maps to locate the places of residence for accused witches (total named accused witches: 3141)
P2 Sur la piste des œuvres antiques (2019)	Sale catalogue and various archival sources of French antiques (including auctioneers' archives), and information on works, sellers, and buyers	SNA; HGI; Visualization	Visualize the biography (origin, career, price trends), the formation and reception of antique objects, and the activities of collectors, institutions, dealers, and intermediaries
P3 Close-Up Cloud (2019)	A collection of historical glass negatives of the collections	Digitalization (Scanning, filming); Visualization	Provide the close-ups of the collections, and enable to examine a wealth of details of individual glass negatives and their depth of field
P4 Dong Qichang Digital Humanities Project (2019)	Dong Qichang's paintings, calligraphies, high-definition images, work texts and treatises	Digitalization; Text Annotation; SNA; HGI; Machine Learning; Visualization	It allows you to visually draw a chronology of Dong Qichang's major works, a network of masters, and so on
P5 Gaoqian Digital Memory Website (2020)	Architectural survey data, oral interview data, audio of family training, photos of folklore scenes, etc. (digital resources totalling 246GB)	Digitalization (Scanning, filming); SNA; 3D Modelling; Text Analysis; Video Production; Chronological Analysis; Visualization	A digital living village has been constructed with a timeline of more than 1,000 years of evolution of the ancient village
.....			
P34 Map-inc: Mapping Incunabula in Greek libraries (2023)	The information of Incunabula (or incunables, which are books printed in Europe before 1501)	Text Annotation; HGI; Visualization	It provides the digital map of the information of Incunabula

Table 1. Content analysis of the 34 selected cases (partial)

Discussion

In this study, we aim to underscore the dual impact—encompassing both contributions and limitations—of DH approaches on historical and cultural resource development. By adapting the input-process-output (IPO) model from systems theory into the source-transformation-result (STR) framework, we systematically analyzed how raw materials (Source) are transformed through technologies and techniques (transformation) into meaningful outcomes (Result). The STR framework facilitates a comprehensive evaluation of how DH approaches influence each stage of historical and cultural resource development, thereby enabling a nuanced understanding of their overall impact.

Contributions and benefits

DH approaches transcend mere technical advancements, embodying a paradigm shift in how we perceive, analyse, and engage with historical and cultural resources. By analysing DH projects through the STR framework, we elucidate several key advantages these approaches offer:

Firstly, at the source end, DH methodologies integrate data from multiple sources, facilitating the regeneration of rich, multifaceted knowledge. This shift from individual to collective analysis enables sophisticated techniques such as cross-textual comparisons and thematic studies, deepening our understanding of historical and cultural dynamics (Champion, 2016; Pacheco, 2022). Additionally, collective analysis dismantles traditional resource silos, promoting a more inclusive and comprehensive historical narrative.

Secondly, the computational techniques central to DH approaches, such as remote reading, empower scholars to analyse vast datasets with unprecedented scale and precision. Utilizing these techniques allows for identifying historical trends, simulating cultural shifts, and reconstructing incomplete datasets, thereby providing new perspectives and insights into previously inaccessible historical phenomena.

Thirdly, interactive, and visual presentations revolutionize the experience and comprehension of history and culture. Tools like dynamic maps, timelines, network diagrams, and 3D models enhance user engagement by enabling interactive exploration of complex historical and cultural contexts. These visualizations transform data presentation into immersive narratives and offer novel interpretations of original resources.

Limitations and challenges

While DH approaches offer numerous benefits for the development of historical and cultural resources, they also bring about some limitations and challenges:

Firstly, the deconstruction and reorganization of corpora through DH methodologies may strip away historical and cultural resources' unique details and contextual richness. For instance, analysing historical letters or diaries for thematic frequency may result in losing emotional depth and personal context embedded within individual documents. The shift from a nuanced perspective to a generalized one can lead to homogenization, overshadowing the rich diversity of cultural expressions and the voices of minority communities. Such processes raise ethical concerns about representation and the preservation of cultural heritage, as the unique stories of underrepresented groups may be subsumed within dominant narratives.

Secondly, although computational technologies and techniques help generate new knowledge due to their unpredictable behavior when handling large datasets (Rodighiero, Derry, Duhaime, et al., 2022), they have limitations regarding depth of interpretation. Historical and cultural resource development often requires understanding subtleties like linguistic style, cultural references, and historical context — areas where computational tools are not yet adept. This limitation is particularly pronounced when dealing with non-English or indigenous texts, where cultural nuances and linguistic intricacies are critical for meaningful interpretation.

Thirdly, while interactive visual representations enhance user engagement, they may oversimplify complex data for ease of use, leading to reductive interpretations of historical and cultural narratives. Visual tools often prioritize clarity and accessibility, which can result in omitting critical details that are less visually striking or harder to quantify. Furthermore, the emphasis on visual engagement may inadvertently prioritize form over substance, where the aesthetic appeal of visualizations takes precedence over their analytical rigor. This challenge underscores the importance of balancing accessibility with accuracy, ensuring that interactive tools serve as gateways to deeper exploration rather than mere endpoints.

Conclusion

The convergence of DH approaches with traditional humanities scholarship presents an unprecedented opportunity for generating new insights into historical and cultural resources. However, it must be pursued cautiously and ethically to ensure that technological advancements do not overshadow the human elements central to historical and cultural studies. In conclusion, finding a balanced approach that optimizes the advantages of DH methodologies while addressing their limitations is vital. Achieving this balance requires interdisciplinary collaboration, fostering dialogue between technologists and humanities scholars to develop hybrid DH tools that blend computational power with humanistic insight. Furthermore, it is crucial to cultivate a reflective DH practice that not only leverages technologies and techniques for resource development but also critically examines the assumptions and biases embedded within these technologies and techniques.

Ultimately, the future of DH lies in its ability to evolve alongside technological progress and humanistic inquiry. By maintaining a commitment to ethical considerations and preserving the integrity of historical and cultural narratives, DH approaches will continue to shape how we engage with the past, understand the present, and imagine the future. Through this ongoing evolution, DH approaches can fulfil their potential as a transformative force in preserving and enhancing human culture and history.

Acknowledgements

We would like to express our sincere gratitude to the National Social Science Foundation for their generous financial support and to the reviewers for their meticulous review and invaluable suggestions.

About the authors

Sijia Lu is a PhD Candidate in the School of Information Resource Management at Renmin University of China, Beijing, China. Her research interests are in digital humanities and scholarly communication. She can be contacted at lu.sijia00@gmail.com.

Jiaqing Long is a PhD Candidate in the School of Information Resource Management at Renmin University of China, Beijing, China. His research interests are in archival science and digital humanities. He can be contacted at jiaqing.long@ruc.edu.cn.

Yongjun Xu is a professor in the School of Information Resource Management at Renmin University of China, Beijing, China. His research interests are in archival science and scholarly communication. He can be contacted at rucxyj@163.com.

References

Berry, D. M., & Fagerjord, A. (2017). Digital humanities: Knowledge and critique in a digital age. John Wiley & Sons.

Champion, E. M. (2016). Digital humanities is text heavy, visualization light, and simulation poor. *Digital Scholarship in the Humanities*, 32(suppl_1), 25-32.

Digital Humanities Awards. (n.d.). *Home page*. http://dhawards.org/

Dobson, J. E. (2019). *Critical digital humanities: The search for a methodology*. University of Illinois Press.

Hendery, R., Kasra, M., Licastro, A., et al. (2019). XR in DH: Extended reality in the digital humanities. *Digital Humanities conference* 2019. https://sites.duke.edu/vardhi/files/2021/03/XR-in-DH-Extended-Reality-in-the-Digital-Humanities.pdf

Huang, Y., & Lu, R. (2015). Research on the classification of historical and cultural resources in China. *Proceedings of the 2nd Annual International Conference on Social Science and Contemporary Humanity Development*, 100-106. Atlantis Press. https://doi.org/10.2991/sschd-16.2016.20

Ilgen, D.R., Hollenbeck, J.R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology*, 56, 517–543.

Karrouche, N. (2019). Still waters run deep: Including minority voices in the oral history archive through digital practice (version 2.0). DataverseNL. https://doi.org/10.34894/T0LLKS

Moretti, F. (2013). *Distant reading*. Verso Books.

Pacheco, A. (2022). Digital humanities or humanities in digital: revisiting scholarly primitives. *Digital Scholarship in the Humanities*, 37(4), 1128-1140.

Rodighiero, D., Derry, L., Duhaime, D., et al. (2022). Surprise machines. Information Design Journal, 27(1), 21–34.

Rodríguez Díaz, Karines, Haber Guerra, Yamile, & Gómez Masjuán, Miguel Ernesto. (2021). Digital Humanities: an approach to emerging methodologies from the case #ElectionsUSA. *Alcance*, 10(25), 301-323.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Systematically modeling and extracting bibliographic metadata of power grid standard documents with LLMs

Guowei Chen, Wei Xie, Yanan Liu, Xiaoqun Yuan, and Liang Zhao

DOI: https://doi.org/10.47989/ir30iConf47233

Abstract

Introduction. This study addresses the critical need for systematic bibliographic metadata representation and extraction from power grid standard documents, essential for operational efficiency and knowledge management in the power industry.

Method. We developed a two-stage methodology utilizing large language models (LLMs) for extracting bibliographic metadata. The first stage involves constructing state grid-oriented instructions for the LLM, and the second stage includes a trustworthiness estimation to ensure the reliability of the extracted metadata.

Analysis. Experiments were conducted using 96 state grid PDF samples to test the accuracy of metadata extraction. The performance of different LLMs was evaluated using single and multiple instructions.

Results. The results showed over 70% accuracy across all models, with GPT-4 achieving the highest accuracy of 84%. Multiple instructions outperformed single instructions, highlighting the effectiveness of our approach.

Conclusion(s). This study demonstrates the promising potential by LLM for data management in the power grid field, with the trustworthiness estimation mechanism significantly enhancing the reliability of the data extracted.

Introduction

Power grid standard documents in actual production provides a set of standardized operation and management processes for power grid enterprises (China Power, 2024). Through unified regulations, standardization helps optimize resource allocation, risk control and decision-making support, thus enhancing overall operational efficiency and service quality (Gal & Rubinfeld, 2019; Reyes et al.,2023).

In the era of burgeoning data and information systems, metadata plays a pivotal role in organizing, navigating, and understanding the plethora of massive standard documentation resources (Zeng & Qin, 2020; Riley, 2017; Baca, 2016). Particularly, bibliographic metadata of the standard documents, which includes information about documents such as titles, authors, and publication details, is crucial for effective knowledge management and retrieval (Liu et al.,2022; Reyes et al.,2023).

However, systematically and effectively extract structured bibliographic metadata from a variety of unstructured standard documents (e.g., in PDF format) is a non-trivial task. Existing methods of knowledge extraction from PDFs (Büchter et al., 2020) typically uses OCR technique to convert these documents into editable text (He ,2020; Karthick et al.,2019;Bartz et al.,2017), followed by manual rule-based systems (Proctor et al.,2019) (e.g., by regular expressions; Chapman et al., 2017) or the construction of machine learning models (Fanni et al.,2023; Kang et al., 2020; Chowdhary& Chowdhary, 2020) to recognize the required information. Such methods generally incur high costs and require model readjustment when the set of standard documents change, lacking a convenient and universal framework.

Ovbiously, there are two major challenges that need addressing. Conceptually, a unified and structured representation framework of bibliographic metadata specifically for power grid standard documents is still ambiguous. Technically, a unified and low-cost knowledge extraction pipeline needs to be developed.

The advent of large language models (LLMs) has marked a significant leap forward in the field of knowledge extraction, especially in areas where traditional data labelling is limited or costly. Efforts have been made to harness the abilities of LLMs for generative information extraction tasks, (e.g. Xu et al., 2023; Sivarajkumar et al., 2023). Researchers have been investigating the usage of LLMs in both zero-shot and in-context learning settings to tackle the problem of extracting procedures from unstructured text in an incremental question-answering fashion (Dagdelen et al., 2024; Kabongo & D'Souza, 2024). For example, Papaluca et al. (2023) developed a dynamic pipeline for knowledge graph triplet extraction in zero-shot and few-shot scenarios by using contextual data from a knowledge base. Xue et al. (2024) introduced AutoRE, a method for relationship extraction that goes beyond traditional sentence-level analysis to document-level analysis. Through simple prompt-based strategies, Shao et al. (2023) explored using pre-trained LLMs for extracting astronomical knowledge entities from astrophysical journal articles. Zhang et al. (2024) proposed a zero-shot learning framework to enable LLMs to assimilate knowledge even without direct training on specific datasets.

Inspired by the promising capabilities of LLMs, in this study, we propose a comprehensive pipeline for both modelling and extracting the bibliographic metadata of power grid standard documents, marking a pivotal step from conceptual representation to application realization within the domain of power grid standard documentation. As shown in Figure 1, we adopted a concise and efficient pipeline to realize metadata extraction from power network standard documents, which consists of two modules, addressing the two major challenges, respectively. Firstly, we conceptually construct a unified and structured framework of bibliographic metadata specifically for power grid standard documents, a step that is the basis of the entire process and involves determining what key information should be extracted from the documents. Secondly, a two-stage methodology

utilizing LLMs for extracting bibliographic metadata is designed, including instruction design upon LLMs and trustworthiness estimation to refine the reliability.

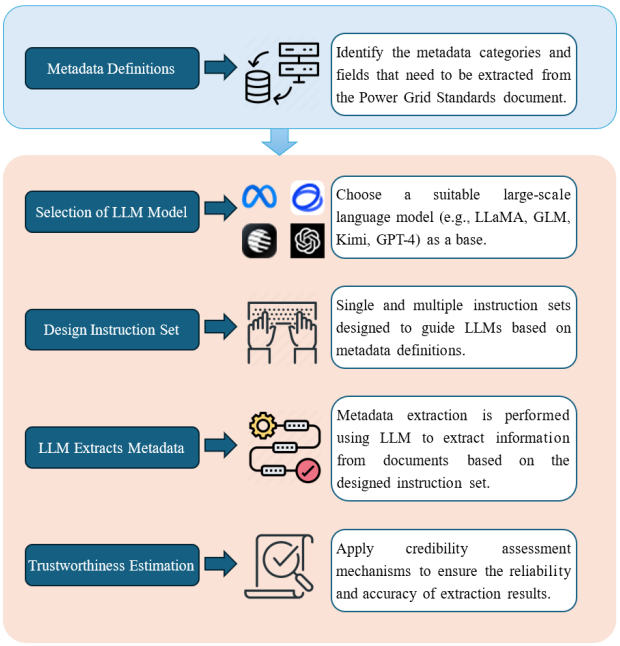


Figure 1. Research framework

Modeling bibliographic metadata of power grid standard documents

Conceptualization of bibliographic metadata

It is necessary to construct a unified and structured conceptualizaiton that categorizes metadata into distinct yet interconnected components. According to the national standard GB/T 22373-2021 'Standard Document Metadata' (http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=174FB61306FB900BFD86FF84C3BD12F7) of China, which is designed to standardize the metadata of general standard documents (Liu et al.,2022), we summarize and design the conceptual framwork of bibliographic metadata by involving the specific attributes in the field of power grid, capturing the essence of the metadata systematically and hierarchically. Totally, 24 bibliographic metadata items are conceptualized, organizd in four categories including basic information, semantic information, correlated information, as well as property information. Figure 2 illustrates hierarchical diagram for clarity.

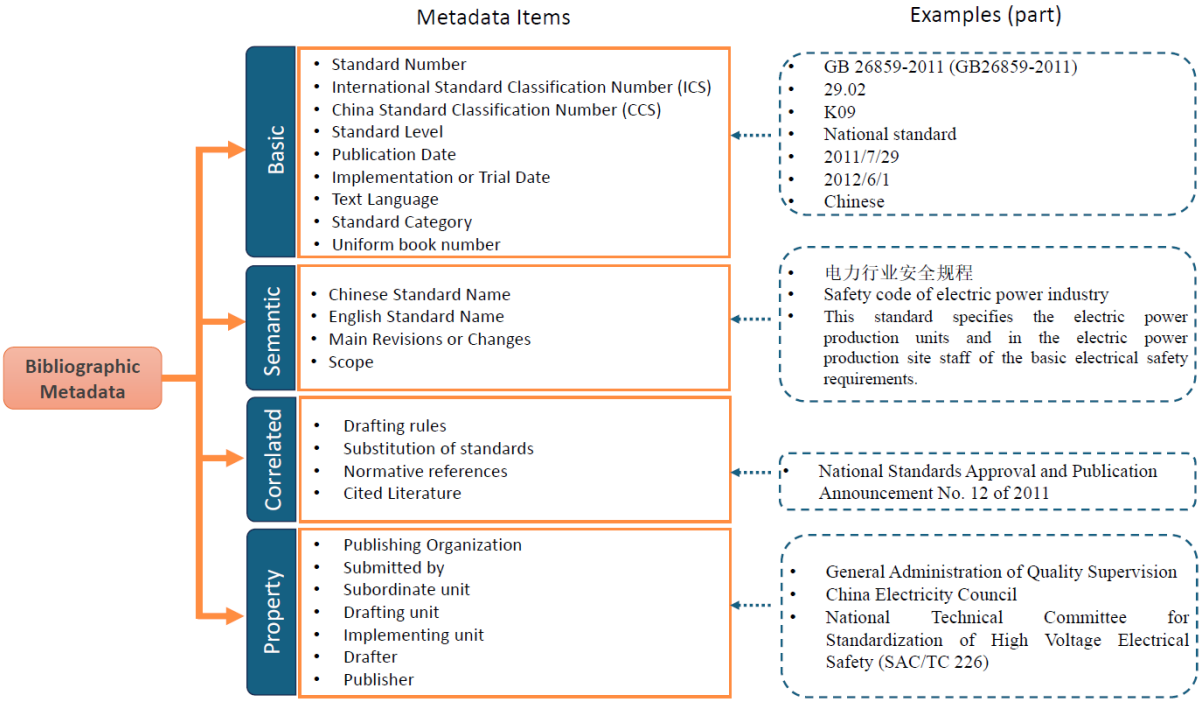


Figure 2. Hierarchical conceptualization of bibliographic metadata

Detailed representations of title metadata

In this section, we will detail the representation of each metadata item in the standard documents according to the above hierarchical conceptualization.

Basic information

Standard number: a character type that consists of the standard code, followed by a space, a sequential number, a hyphen, and a four-digit publication year. This is located on the front page of the standard document.

International standard classification number (ICS): a character type used for international classification, providing a standardized reference for the document's subject matter.

Chinese standard classification number (CCS): a character type that categorizes the standard within the Chinese regulatory framework.

Standard level: indicates the hierarchy of the standard, such as GB (National Standard), DL (Industry Standard), T (Group Standard), Q (Enterprise Standard), and DB (Local Standard).

Release date: presented in YYYY-MM-DD format, indicates when the standard was officially published.

Implementation or trial date: also in YYYY-MM-DD format, specifies when the standard becomes effective or enters a trial phase.

Text language: identifies the language of the document, such as Chinese or English.

Standard category: classifies the standard into types such as terminology, symbols, classification, testing, specifications, codes of practice, and guidelines.

Uniform book number: a character type that serves as a unique identifier for the publication.

Semantic information

This category captures the content-related aspects of the standard, providing insights into its purpose and scope.

Chinese standard name: the official title of the standard in Chinese, presented as free text.

English standard name: the official title of the standard in English, also presented as free text.

Main revisions or changes: a free text description of significant revisions or changes made to the standard, with different entries separated by semicolons.

Scope: defines the applicability of the standard, detailing what is covered and the contexts in which it is relevant.

Correlated information

This category includes references and relationships that connect the standard to other documents and frameworks.

Drafting rules: refers to the standards that guided the drafting process, indicated by their standard numbers. This information is typically found in the preface of the document.

Replaced standards: lists the numbers and names of standards that this document replaces, with multiple entries separated by semicolons.

Normative reference documents: free text that includes standard numbers and descriptions of documents that are considered normative, with multiple items separated by semicolons.

Cited literature: free text that may include standard numbers and names of cited documents, with multiple entries separated by semicolons. This is typically found on the last page and copyright page of the document.

Property information

This category encompasses details regarding the ownership and responsibility for the standard.

Publishing organization: the entity responsible for publishing the standard, ensuring its formal distribution and recognition, typically presented as free text.

Submitted by: the organization that submitted the standard for review and publication, which may include government bodies or industry groups, with multiple entities separated by commas.

Subordinate unit: any subsidiary or subordinate units under the main organization that contributed to the standard, listed in free text.

Drafting unit: the specific group or department within an organization tasked with drafting the standard document, presented as free text.

Implementing unit: the entity responsible for the enforcement and practical application of the standard, detailed in free text.

Drafter(s): individuals involved in drafting the standard, with multiple names separated by commas to acknowledge all contributors.

Publisher: the organization or entity that physically produces and disseminates the final version of the standard document, typically indicated as free text.

Prompt-based extracting bibliographic metadata by LLMs

Taking the abilities of LLMs, we design a two-stage prompt-strategy to extract the bibliographic metadata from the standards documents in a simple-yet-powerful zero-code manner.

Stage 1: state grid-oriented instruction construct

The first stage of our method is the '*State Grid-Oriented Instruction Construct*', where we meticulously design and implement a set of instructions aimed at guiding the Large Language Model (LLM) in extracting metadata from power grid standard documents. This stage is crucial as it sets the groundwork for how the LLM interacts with the documents and identifies the necessary metadata elements. Through these instructions, we effectively prompt the LLM to perform the task of metadata extraction, ensuring that it focuses on the relevant details and follows a structured approach to gather the required information.

Identity instruction

This process is to prompt the LLM as a meta-data extraction expert. Then tell the LLM what to do next. The instruction is as follows: *now you are a metadata extraction expert. I will provide the PDF to be extracted and the extracted metadata items. Please help me extract the relevant content.*

Metadata instruction

Two alternative prompting strategies are designed to extract the metadata.

Multiple instructions (MI) extract items based on the different locations of metadata within the standard document.

1) Prompts for metadata items in the homepage of standard files:

> *please output the following content in sequence: Standard number, International standard classification number, Chinese standard classification number, Standard level (GL-national standard, DL-industry standard, T-group standard, Q-enterprise standard, DB-local standard), Chinese standard name, English standard name, Release date, Implementation or trial date, Release organization, Text language, Standard category.*

2) For metadata items in the preface of standard files:

> *please output the following content in order: Drafting rules, Replacement standards, Main revisions or changes, Proposing unit, Responsible unit, Drafting unit, Implementing unit, Drafter.*

3) For metadata items in the body part:

> *please input the following content in order: Specified scope, Applicable scope, Inapplicable scope, Normative reference documents (multiple items separated by ;).*

4) For metadata items in the copyright page of standard files:

> *Please input the following content in order: References (multiple items separated by ;) , Publisher, Uniform book number.*

Single instruction (SI) simply extracts all metadata without location hints as following:

> *output the following content in sequence: Standard number, International standard classification number, Chinese standard classification number, Standard level (GL-national standard, DL-industry standard, T-group standard, Q-enterprise standard, DB-local standard), Chinese standard name, English standard name, Release date, Implementation or trial date, Release organization, Text language, Standard category, Drafting rules, Replacement standards, Main revisions or changes, Proposing unit, Responsible unit, Drafting unit, Implementing unit, Drafter, Specified scope, Applicable scope, Inapplicable scope, Normative reference documents (multiple items*

separated by ;), Please input the following content in order: References (multiple items separated by ;) , Publisher, Uniform book number.

Stage 2: trustworthiness estimation (TE)

Following the metadata extraction facilitated by the State Grid-Oriented Instructions, the second stage of our method is the '*Trustworthiness Estimation.*' In this phase, we evaluate the credibility of the metadata extracted by the LLM to enhance the reliability of the model's results. This involves assessing the consistency and accuracy of the extracted data, ensuring that the final output is not only comprehensive but also trustworthy. By implementing a trustworthiness estimation mechanism, we aim to mitigate the inherent variability in LLM responses and select the most reliable metadata, thereby ensuring the highest quality of information for our users.

It is a common phenomenon that LLM's Randomness, i.e. if you prompt an LLM with the same instruction multiple times, it may give different answers. To overcome this challenge, we propose a trustworthiness estimation mechanism to make our instruction method more reliable.

We using the instructions to prompt LLM k times. For each unique meta-data answer, we define its trustworthiness as:

$$t = \frac{h}{k}$$

where h denotes the appearing time of this answer. Lastly, we choose the most trustworthy i-th answer as final answer, where

$$i = \text{argmax}(T(i)).$$

Experimental study and result

Experiment setting

Dataset: we conducted the experiment on 96 state grid PDF samples. All these 96 samples are treated as testing data. As the LLMs do not need to be trained or fine-tuned.

Evaluation metric: we use accuracy as the evaluation metric. For each sample, the accuracy is calculated as the ratio of right metadata items in total items. The final results are defined as the average performance on all testing samples.

Baselines: four baselines are considered, GPT-4, Kimi, GLM, and LLaMA (7B).

Settings: we directly use the official API of each LLM to conduct our experiment.

Result

Main results

Our extensive testing of various LLMs using both single and multiple instruction methods (with TE by default) has yielded promising results shown in Table 1. All models have achieved an average accuracy of over 70%, validating the effectiveness of our instructional approach in extracting metadata. Notably, the GPT-4 model has demonstrated exceptional performance with the highest accuracy of 84% by multiple instructions, affirming its robustness as a backbone model for the task. This superior performance is likely due to GPT-4's advanced natural language processing capabilities, which allow it to better comprehend and respond to complex instructions.

Moreover, a consistent trend emerges, revealing that multiple instructions consistently outperform single instruction across all models. This trend suggests that the additional context provided by multiple instructions enables the models to generate more precise and detailed responses, thereby enhancing the overall accuracy of metadata extraction.

	LLaMA	GLM	Kimi	GPT-4
SI without TE	0.68	0.72	0.69	0.73
SI	0.72	0.75	0.74	0.77
MI without TE	0.73	0.77	0.76	0.82
MI	0.79	0.82	0.82	0.84

Table 1. Collation of experimental results

Comparisons between multiple and single instructions

Figure 3 give a case study comparing the differences between the two instructions. This visual representation vividly captures the nuances in the responses generated by each method. Single instruction, while efficient for basic tasks, tend to yield only abbreviated metadata, providing a quick snapshot but lacking in depth. Conversely, multiple instructions, which are more elaborate and structured, result in a richer output that includes both the abbreviated forms and the full names of metadata elements. This comprehensive approach not only enhances the granularity of the data but also significantly improves the contextual understanding of the metadata, thereby facilitating more informed decision-making processes.

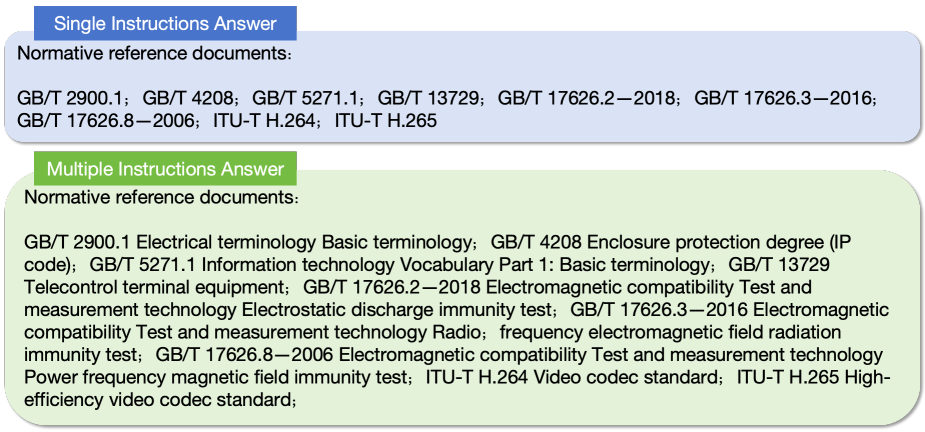


Figure 3. Comparison of metadata extraction responses: single vs. multiple instructions

This observation underscores a critical point that the depth and breadth of instructions have a direct impact on the quality of metadata extraction. By leveraging multiple instructions, we are essentially equipping the LLMs with a more detailed roadmap, enabling them to navigate the complexities of document structures and extract metadata that is not just accurate but also contextually rich.

Ablation results

To further validate the significance of our Trustworthiness Estimation (TE) mechanism, we conducted an ablation study. The results, as indicated in the table, reveal a notable decrease in performance when the TE mechanism is not utilized. This decline underscores the pivotal role of the TE mechanism in bolstering the reliability of the LLMs. By incorporating TE, we are effectively mitigating the inherent variability in LLM responses, ensuring that the extracted metadata is not only accurate but also consistent across different instances.

Discussion and conclusion

Our study presents a comprehensive framework tailored for the power system industry, focusing on the systematic extraction of bibliographic metadata from standard documents with LLMs, contributing a significant advancement of document management and knowledge extraction in vertical fields like power grid.

Theoretical and practical implications

The theoretical underpinning of this research is profound, marking a pioneering effort in applying LLMs to metadata extraction within the power industry. We propose the first trial of a comprehensive and unified conceptualization of bibliographic metadata for powergrid standard documents. On the other hand, the exploration of the simple-yet-powerful prompt-based strategy with zero code manner in our study transcends the conventional boundaries of AI application, offering a fresh lens through which to view the integration of artificial intelligence with industry-specific processes. By situating our study within the operational context of power systems, we contribute to the broader discourse on AI's role in enhancing operational efficiency and data integrity.

From a practical standpoint, this study underscores the transformative potential of LLMs in streamlining operational workflows and bolstering data management practices within the power industry (Schilling-Wilhelmi et al.,2024; De Santis et al.,2024). Compared with traditional knowledge extraction methods with high cost and complicated realization, such as regular expressions (RE; Chapman et al., 2017) and machine learning models (Fanni et al.,2023; Kang et al., 2020; Chowdhary& Chowdhary, 2020), we demonstrate a universal, convenient but effective pipeline with LLMs. The ability of LLMs to parse through complex documents and extract critical metadata with high accuracy not only enhances the speed of information retrieval but also mitigates the risk of human error. In addition, it also provides a new but more user-friendly interaction mode with natural language, instead of complicated RE grammar rules and machine learning algorithms traditionally leveraged in metadata extraction.

Limitations and future work

Of course, there are also limitations in our study. One of the limitation of our current study is the modest size of the dataset involved in the experiments, which, while representative of standard documents in the power industry, may not encapsulate the full diversity of documents encountered in real-world applications. In the future work we will continue to investigate the feasibility and robustness of our model by the expansion of dataset to encompass a more comprehensive and heterogeneous collection of documents. Additionally, as we only explored the simple zero-shot prompt strategy to interact with LLMs, the model's performance, while basically satisfactory, may not fully account for the complexities and variability in language use across different documents and contents. Future work will investigate the prompts with few-shots to help LLMs better understand the tasks and also explore the integration of the retrieval-augmented generation (RAG) model to further improve the accuracy in metadata extraction.

Acknowledgements

This work was supported by State Grid Corporation of China (Comprehensive Construction Technology of Standard Text Resources and Key Elements, 5400-202318585A-3-2-ZN), and Key Laboratory of Semantic Publishing and Knowledge Service of the National Press and Publication Administration (Wuhan University).

About the authors

Guowei Chen works at State Grid Fujian Electric Power Co.,Ltd, Fuzhou, China. His research interests are technology management and knowledge services, and can be contacted at chen_guowei@fj.sgcc.com.cn

Wei Xie works at State Grid Fujian Electric Power Research Institute, Fuzhou, China. His research interest is artificial intelligence in electric power and can be contacted at xiewei3896@163.com

Yanan Liu is an Editor at Yingda Media Investment Group Co., Ltd. Her research interest is digitalization of electric power standards, and she can be contacted at 743113090@qq.com.

Xiaoqun Yuan is Associate Professor in School of Information Management, Wuhan University, China. He received their Ph.D. from Huazhong University of Science and Technology. His research interests are information resource management and knowledge services and can be contacted at yuan20030308@whu.edu.cn

Liang Zhao is Associate Professor in School of Information Mangement, Wuhan University, China. She received her Ph.D. from Tsinghua University. Her research interests are data mining and knowledge services. She is the corresponding author of the paper and can be contacted at liangzhao@whu.edu.cn

References

Baca, M. (Ed.). (2016). *Introduction to metadata*. Getty Publications.

Bartz, C., Yang, H., & Meinel, C. (2017). STN-OCR: A single neural network for text detection and text recognition. *arxiv preprint arxiv*:1707.08831.10.48550/arXiv.1707.08831

Büchter, R. B., Weise, A., & Pieper, D. (2020). Development, testing and use of data extraction forms in systematic reviews: a review of methodological guidance. BMC medical research methodology, 20, 1-14. 10.1186/s12874-020-01143-3

Chapman, C., Wang, P., & Stolee, K. T. (2017). Exploring regular expression comprehension. In 2017 *32nd IEEE/ACM International Conference on Automated Software Engineering* (ASE), 405-416. 10.5555/3155562.3155616

China Power. (2024, July 19). State Grid lays a solid foundation for the construction of a new type of power system.

Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649. 10.1007/978-81-322-3972-7

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ... & Jain, A. (2024). Structured information extraction from scientific text with large language models. Nature Communications, 15(1), 1418. 10.1038/s41467-024-45563-x

De Santis, A., Balduini, M., De Santis, F., Proia, A., Leo, A., Brambilla, M., & Della Valle, E. (2024). Integrating Large Language Models and Knowledge Graphs for Extraction and Validation of Textual Test Data. *arxiv preprint arxiv*:2408.01700. 10.1007/978-3-031-77847-6

Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). Natural language processing. In *Introduction to Artificial Intelligence* (pp. 87-99). Cham: Springer International Publishing. 10.1007/978-3-031-25928-9

Gal, M. S., & Rubinfeld, D. L. (2019). Data standardization. *NYUL Rev.*, 94, 737.

He, Y. (2020). Research on text detection and recognition based on OCR recognition technology. In 2020 *IEEE 3rd International Conference on Information Systems and Computer Aided Education* (ICISCAE), 132-140. 10.1109/ICISCAE51034.2020.9236870

Kabongo, S., & D'Souza, J. (2024). Instruction Finetuning for Leaderboard Generation from Empirical AI Research. arxiv preprint arxiv:2408.10141. 10.48550/arXiv.2408.10141

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, 7(2), 139-172. 10.1080/23270012.2020.1756939

Kapoor, A., Gulli, A., Pal, S., & Chollet, F. (2022). *Deep Learning with TensorFlow and Keras: Build and deploy supervised, unsupervised, deep, and reinforcement learning models*. Packt Publishing Ltd.

Karthick, K., Ravindrakumar, K. B., Francis, R., & Ilankannan, S. (2019). Steps involved in text recognition and recent research in OCR; a study. *International Journal of Recent Technology and Engineering*, 8(1), 2277-3878.

Liu, J., Kong, Y., & Peng, G. (2022, June). Interpreting the Development of Information Security Industry from Standards. In *International Conference on Human-Computer Interaction* (pp. 372-391). Cham: Springer International Publishing. 10.1007/978-3-031-05463-1

Papaluca, A., Krefl, D., Rodriguez, S. M., Lensky, A., & Suominen, H. (2023). Zero-and few-shots knowledge graph triplet extraction with large language models. arxiv preprint arxiv:2312.01954. 10.48550/arXiv.2312.01954

Proctor, M., Fusco, M., Vacchi, E., & Sottara, D. (2019). Rule modularity and execution control enhancements for a Java-based rule engine. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering* (AIKE) , 89-96. 10.1109/AIKE.2019.00023

Reyes, J., Mula, J., & Díaz-Madroñero, M. (2023). Development of a conceptual model for lean supply chain planning in industry 4.0: multidimensional analysis for operations management. *Production Planning & Control*, 34(12), 1209-1224. 10.1080/09537287.2021.1993373

Riley, J. (2017). Understanding metadata. *Washington DC, United States: National Information Standards Organization* (http://www.niso.org/publications/press/UnderstandingMetadata.pdf), 23, 7-10.

Schilling-Wilhelmi, M., Ríos-García, M., Shabih, S., Gil, M. V., Miret, S., Koch, C. T., ... & Jablonka, K. M. (2024). From Text to Insight: Large Language Models for Materials Science Data Extraction. *arxiv preprint arxiv:*2407.16867. 10.48550/arXiv.2407.16867

Shao, W., Zhang, R., Ji, P., Fan, D., Hu, Y., Yan, X., ... & Chen, L. (2024). Astronomical knowledge entity extraction in astrophysics journal articles via large language models. Research in Astronomy and Astrophysics, 24(6), 065012.

Singh, P., & Manure, A. (2019). *Learn TensorFlow 2.0: Implement Machine Learning and Deep Learning Models with Python*. Apress.

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2024). An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Medical Informatics, 12, e55318. 10.2196/55318

Stevens, E., Antiga, L., & Viehmann, T. (2020). *Deep learning with PyTorch*. Manning Publications.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., ... & Chen, E. (2023). Large language models for generative information extraction: a survey. arxiv preprint arxiv:2312.17617. 10.1007/s11704-024-40555-y

Xue, L., Zhang, D., Dong, Y., & Tang, J. (2024). AutoRE: Document-Level Relation Extraction with Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for

Computational Linguistics (Volume 3: System Demonstrations) , 211-220. 10.48550/arXiv.2403.14888

Zeng, M. L., & Qin, J. (2020). *Metadata*. American Library Association.

Zhang, J., Ullah, N., & Babbar, R. (2024). Zero-shot learning over large output spaces: utilizing indirect knowledge extraction from large language models. arxiv preprint arxiv:2406.09288.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Demystifying perception and decision-making in peer review: a semantic perspective

Yujie Zhang, Weikang Yuan, and Zhuoren Jiang

DOI: https://doi.org/10.47989/ir30iConf47230

Abstract

Introduction. Peer review is fundamental to scientific progress. However, it remains a *'black box'* due to the ambiguous and subjective decision-making rationale, hidden within reviewers' perceptions and judgments.

Research gaps. The scarcity of peer review data and the limited analytical scope have hindered empirical studies from thoroughly exploring these underlying factors.

Method. We collected 114,006 review records from OpenReview. Developing deep learning models, we extracted reviewers' perceptions of key factors from the review content.

Analysis. Our analysis focuses on how these perceptual factors influence reviewers' decisions and outcomes of the papers.

Results. Originality is the element that reviewers value and recognize the most, followed by motivation. In contrast, chairs prioritize the overall validity of the work within the field, giving more weight to the substance and soundness of papers. The role of motivation is generally supportive but has limited distinguishing power on final decisions. Perceived clarity, replicability, and meaningful comparison have a relatively minor role in the review process.

Conclusion. Our study highlights the decision-making process based on reviewers' perceptions and uncovers perceptual differences between reviewers and chairs. This contributes to greater transparency in the peer review and helps build trust within the scientific community.

Introduction

Peer review is considered the backbone of scientific construction (Alberts et al., 2008). By filtering out substandard work and providing expert feedback, it helps ensure the quality and credibility of scientific research, contributing to the evolution and enrichment of knowledge (Bailly, 2016; Jefferson et al., 2002; Riley & Jones, 2016). However, despite its importance and decisiveness, the decision-making process in peer review mechanisms often resembles a black box (Chong & Mason, 2021; Cutler et al., 2022), which may raise concerns within the scientific community. It arises not only from the non-transparency of the procedure (Hamilton et al., 2020) but also from the ambiguous and subjective rationale underlying the reviewers' decision-making.

The review decision-making process involves complex cognitive processes (Brosch et al., 2013; Lindsay & Norman, 2013; Marietta & Barker, 2019), with perceptions and judgments formed based on the presented facts (Dror & Fraser-Mackenzie, 2008; Street & Ward, 2019). It is complex and subjective, influenced by reviewers' cognitive biases and qualifications (Benda & Engels, 2011). Their decisions—whether to accept or reject a paper—are affected by factors they have perceived such as significance, originality, methodological rigor, and presentation (Blockeel et al., 2017).

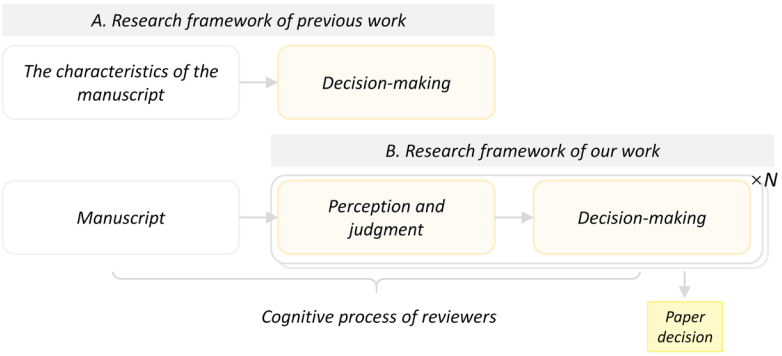


Figure 1. The research framework of previous studies compared to our work. Prior empirical studies on peer review decision-making have focused on extracting the objective characteristics of papers to reveal the preliminary process of reviewers' decisions. However, the reviewers' perceptions and judgments remain unexplored.

Existing research empirically unveils reviewers' decision-making rationale, biases, or preferences, typically employing a '*fact-decision*' approach (Figure 1). For instance, they utilize external metrics, such as the disruption index, to objectively describe the characteristics of papers and then examine whether reviewers' decisions show a preference for certain work (Liang et al., 2023; Teplitskiy et al., 2022). This approach overlooks the crucial intermediate step—the '*perception and judgment*' phase, closely intertwined with the cognitive processes of reviewers (Lamont, 2009; Lee et al., 2013). Exploring this can offer valuable insights into the cognitive aspects of decision-making in peer review, enhance transparency, and build greater trust in the academic community.

Nevertheless, there are still some critical research gaps that may hinder further exploration: (a) The first challenge is about how to measure the reviewer's perception. Generally, reviewers' perceptions could be revealed in their comments like enthusiastic praise or critical criticism (Mavrogenis et al., 2020). However, most peer review processes are not publicly available and related data is extremely scarce (Balietti et al., 2016), making it very difficult to obtain reviewer comments in real-world scenarios. Furthermore, the review contents are usually unstructured data, which makes it harder to extract meaningful insights. Prior studies examined review content through single-dimensional features like sentiment (Matsui et al., 2021), tone (Nobarany & Booth,

2015) and length (Maddi & Miotti, 2024), but have struggled to capture the multifaceted nature of the perceptions. (b) Few empirical studies assess the relationship between reviewers' perceptions and decision-making. This lack of evidence hampers theory development and highlights gaps in understanding the peer review process and perceptual factors of reviewers.

To address these research gaps, we propose the following research questions (RQs):

> **RQ1:** how can we automate the extraction of reviewers' perceptions and judgments of a paper?
>
> **RQ2:** from the perspective of reviewers' decisions, how do perceptual factors influence reviewers?
>
> **RQ3:** from the perspective of papers' decisions, how do reviewers' perceptions affect the decision of a paper?

To answer the above research questions, we collected a substantial number of open review records from OpenReview and carefully filtered them to construct a valuable peer-review dataset, comprising 114,640 records from top-tier computer science conferences. Then, we develop deep learning models to automatically extract and measure reviewers' multi-dimensional perceptual factors (RQ1). Based on these factors, we conceptualize the peer review decision-making process into two primary phases: individual reviewer decisions (RQ2) and paper decisions affected by multiple reviewers (RQ3) and examine the relationship between reviewers' perceptions and decision-making in these phases.

Related work

Perception in peer review has been discussed by existing studies, particularly regarding how reviewers perceive the peer review process itself. Existing research has explored various aspects of reviewers' perceptions, such as their motivations for reviewing, preferred roles, and preferences regarding review mechanisms (Curtin et al., 2018). Other studies have examined reviewers' willingness to receive training, accept feedback, and sign their reviews (Manchikanti et al., 2015). Less work has empirically investigated reviewers' perceptions of the paper. One such study uses surveys to quantify the gap between researchers and reviewers regarding the perceived importance of specific validity and reliability issues (Street & Ward, 2019).

The process from perception to decision in peer review is inherently subjective, and this subjectivity plays a crucial role in shaping the overall review process. Existing work argues that such subjectivity is not only unavoidable but also necessary, as it may help mitigate the herding effect and encourage diverse perspectives (Park et al., 2014). This subjectivity also contributes to the phenomenon of low inter-rater reliability among reviewers. While some researchers view this inconsistency as a natural and inevitable aspect of peer review (Roediger, 1991), others consider it problematic, suggesting that it undermines the reliability of the evaluation process (Marsh et al., 2007).

The process of review is not merely about evaluating the content of a paper; it is also shaped by '*sense of self and their relative positioning*' of reviewers (Lamont, 2009). The connection between identity and evaluation raises concerns about the potential for cognitive biases, which are often cited as a threat to the progress and integrity of a field (Benda & Engels, 2011). Specifically, reviewers may show a preference for submissions from authors with similar '*schools of thought*', a phenomenon known as '*cognitive cronyism*' (Travis & Collins, 1991). Thus, while subjectivity in peer review can foster diverse viewpoints, it also presents challenges that must be addressed to ensure the fairness and accuracy of the evaluation process.

Dataset

To address the issue that the majority of journals and conferences' reviews are not publicly available, open peer review offers a promising solution to enhance the transparency and fairness of the review process (Wolfram et al., 2020). OpenReview (Soergel et al., 2013) is one of the most famous open peer review websites to promote transparency in scientific communication. It provides submitted papers alongside their corresponding reviews and comments. After a thorough examination of each venue in OpenReview, we focused on the top-tier machine learning conferences—ICLR and NeurIPS for our study (ICLR makes reviews public regardless of acceptance, while NeurIPS allows authors of rejected papers to decide whether to make their reviews public). We collected all publicly available review data for these conferences. After data pre-processing and meticulous data filtering, we obtained review information for 30,440 papers from the year 2013 to 2024, comprising 114,640 records.

Measurement of variables

Automatical extraction perceptual factors of reviewers (RQ1)

'*Perceptual factors*' can be defined as the reviewer's subjective perception and judgment elements, which guide their review decision-making. The reviewer's perception of one submission can be directly reflected in the review comments (Mavrogenis et al., 2020). According to previous studies (Yuan et al., 2022), we identify seven key perceptual factors(These dimensions follow the ACL review guidelines): motivation, originality, soundness, substance, replicability, meaningful comparison, and clarity.

To automatically and effectively measure these factors, we designed a pipeline that includes a discriminative model and a scoring model, each based on RoBERTa (Liu, 2019). Specifically, the discriminative model determines whether a sentence is perception related. If it does, the scoring model then assesses the multidimensional perceptual factors. To illustrate these indicators more clearly, we provide an implication and a simplified example in Figure 2. The detailed training process is outlined as follows (The hyperparameters for both models are set as follows: batch size of 16 and a learning rate of 5e-6. We employ the best model among 8 epochs with the lowest validation loss. The model is trained on an NVIDIA V100 32GB GPU):

We use AMPERE (Hua et al., 2019) and ASAP (Yuan et al., 2022) datasets to train our models. To train the discriminative model, we leverage AMPERE, which categorizes review comments into six types. We defined the task as a classification problem to identify whether a sentence belongs to the category '*evaluation*'. To train the scoring model, we use ASAP, which annotated seven dimensions of text spans with sentiment polarity. We converted it into a numeric labeling task. During training, the model takes the text as input and produces a 7-dimensional output, with each dimension representing a dimension. We use the Tanh function to scale each dimension. The loss is computed using Mean Squared Error (MSE). On the validation set (20% of total data), the model achieved a classification accuracy of 0.83 and an absolute deviation of 0.084 for each dimension.

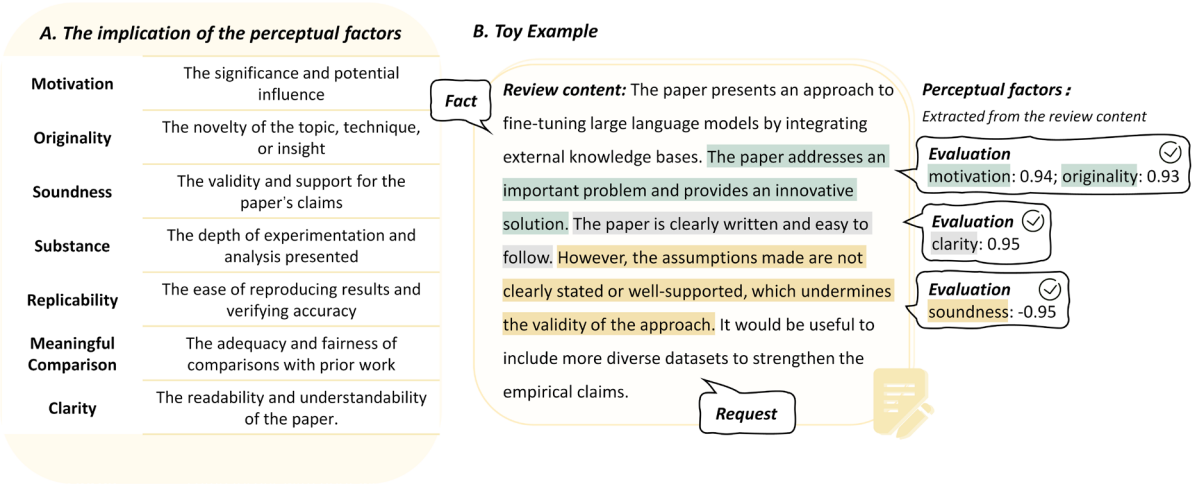


Figure 2. The implication (A) and toy example (B) of the extracted perceptual factors. Each factor's polarity is represented by a continuous value ranging from -1 to 1, where a value closer to 1 indicates higher confidence in a positive attitude. The first and last sentences in the example are categorized as '*fact*' and '*request*', respectively. Unlike the '*evaluation*' category, these sentences exhibit less subjectivity and are thus excluded from measuring the reviewer's perceptual factors. The three sentences in between cover various perceptual factors such as '*motivation*', '*originality*', '*clarity*', and '*soundness*', showcasing the reviewer's diverse perceptions and judgments of the paper.

After measuring the perceptual factors of each sentence, we aggregate the scores as follows:

$$PF_{pi} = \sum PF_{seg}, seg \in T_{pi}$$

$$PF_{p} = \frac{1}{N_p} \sum PF_{pi}$$

Where PF_{pi} represents the 7-dimensional vector of perceptual factors given by the *i-th* reviewer. *seg* represents the sentence of review content T_{pi}. PF_p denotes the average judgment of all reviewers of the paper. The more frequently reviewers mention and emphasize a particular dimension in their comments, the higher the perceptual score for that dimension.

Reviewer decision

Reviewer decisions are a key criterion for chairs in determining paper acceptance, such as '8: *accept, good paper*' or '3: *reject, not good enough*'. Because different conferences use varying scoring ranges from year to year, we standardize all scores to a range of 0 to 10 to ensure comparability. The formula for this calculation is as follows:

$$RD_{pi} = \frac{r_{pi} - min(R)}{max(R) - min(R)} \times 10, r \in R$$

$$RD_{p} = \frac{1}{N_p} \sum \mathrm{RD}_{pi}$$

Where p represents a paper reviewed at a particular conference (held in a specific year). R denotes all the ratings given at this conference. r_{pi} refers to the original rating of the paper p given by the *i-th* reviewer, while RD_{pi} represents the processed one. N_p is the total number of reviews that the paper p received. RD_p is the average rating of all reviewers.

Result

Descriptive statistics of reviewers' perceptual factors

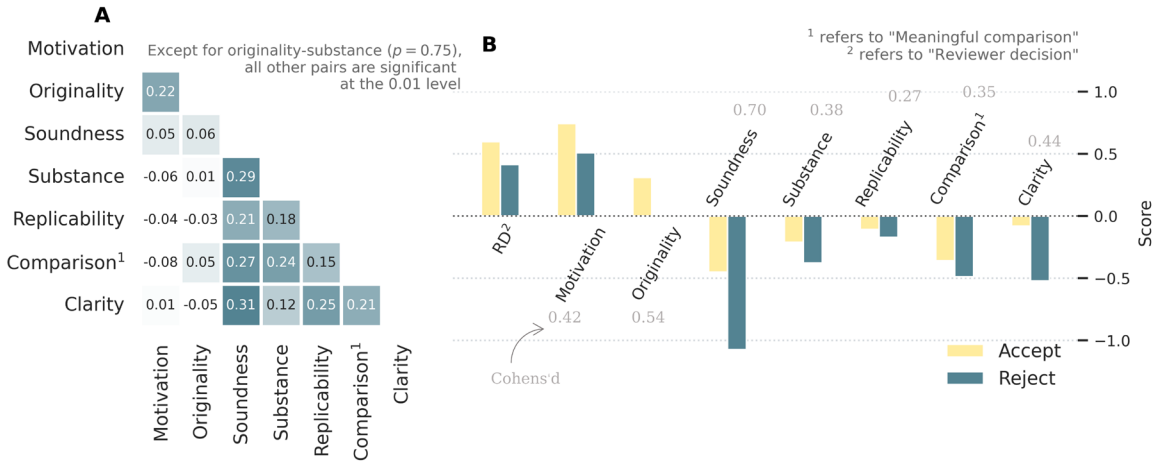


Figure 3. Description of the reviewers' perceptual factors. (A) We calculated the Pearson correlation between different factors using data samples from individual reviews. It is shown that originality and motivation were the most correlated (and vice versa); soundness and clarity were the most correlated (and vice versa); substance and soundness were the most correlated; replicability and clarity were the most correlated; and meaningful comparison and substance were the most correlated. (B) We compared the average of reviewer decisions (*RD*) and perceptual factors for accepted and rejected papers. Unlike Figure 3 (A), which is based on single reviews, the data in Figure 3 (B) are based on individual papers, each with multiple reviews. We found that perceptual factors, whether relating to significance (Motivation), novelty (originality), or rigor (soundness, substance, e.g.), all have the potential to influence the decisions made about the papers. It shows the differences for each factor between the accepted and rejected categories, represented by the Cohen's d.

Firstly, we observed the internal correlation of perceptual factors within reviewers' comments. As shown in Figure 3 (A), we found that the internal correlations of these metrics were generally low (*≤0.31*), which means that there is a lower likelihood of serious multicollinearity issues in modeling.

Figure 3 (B) demonstrates that accepted papers receive higher evaluations from reviewers across all criteria compared to rejected ones. Rejected papers often get positive feedback on motivation but tend to receive negative or no feedback (-0.01) on originality. This indicates reviewers are more supportive of motivation regardless of their recommendation but more critical of originality in papers they dislike. Further, reviewers often give negative evaluations on rigor-related dimensions to prompt authors to improve.

Perception-based decision-making analysis

Reviewer decision analysis (RQ2)

We analyse individual reviewers to examine how their decisions are influenced by their perceptions and judgments. The model is specified as follows:

$$RD_{pi} = \beta_0 + \beta PF_{pi} + \beta' X_{pi} + \alpha_p + \lambda_p$$

RD_{pi} represents the reviewer decision given by the *i-th* reviewer for paper *p*. The key variable PF_{pi} denotes the vector of the reviewer's perceptual factors. X_{pi} is a vector of control variables, including the request number and confidence of the reviewer. The request number represents the number of sentences related to the reviewers' requests. Confidence is the assurance of reviewers, which is scored by themselves. Since the review process is double-blind, we do not control for author-related information in this model. α_i and λ_i represent the fixed effects, capturing stable factors

affecting the reviewer's decision at the venue and time levels respectively. β_0 represents the constant term. We conducted a VIF test (Table S1) and used OLS regression for our analysis (Table 1).

We identified the impact of various perceptual factors on reviewer decisions. Among all perceptual factors (Figure 4), reviewers prioritize originality, followed by motivation, as the primary basis for their decisions. Next in importance are soundness and substance, underscoring the paper's overall persuasiveness and rigor. Lastly, reviewers focus on meaningful comparison, clarity, and replicability. They determine whether a paper is easy to follow and disseminate and they also form the basis for the paper's credibility.

In Figure 5, we conducted a heterogeneity analysis of reviewers grouped by high and low confidence levels (the boundary is set at 5). The results showed that decisions made by high-confidence reviewers were more strongly associated with perceptual factors (except for replicability and clarity).

Dependent Variable: Reviewer decision	Coefficient	SE	P value	Confidence interval
Motivation	0.244	0.005	0.000	[0.235, 0.253]
Originality	0.372	0.005	0.000	[0.363, 0.381]
Soundness	0.214	0.003	0.000	[0.208, 0.220]
Substance	0.213	0.006	0.000	[0.201, 0.225]
Replicability	0.098	0.011	0.000	[0.076, 0.120]
Meaningful comparison	0.172	0.007	0.000	[0.158, 0.187]
Clarity	0.159	0.003	0.000	[0.154, 0.165]
Request number	0.020	0.001	0.000	[0.018, 0.021]
Confidence	-0.060	0.002	0.000	[-0.065, -0.056]

Table 1. The regression result of reviewer decision analysis. The model has an R-squared value of 0.27, and the observation number is 107,272.

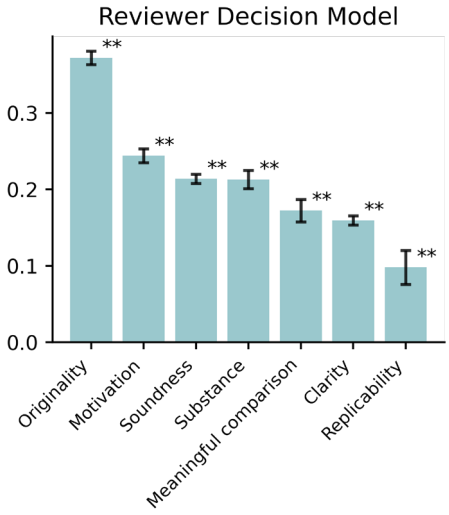


Figure 4. The coefficient comparison of reviewer decision analysis. Among all the perceptual factors, originality has the greatest impact on the reviewer's decision, followed by the perception and judgment of motivation. ** represents $p < 0.01$, and * represents $p < 0.05$.

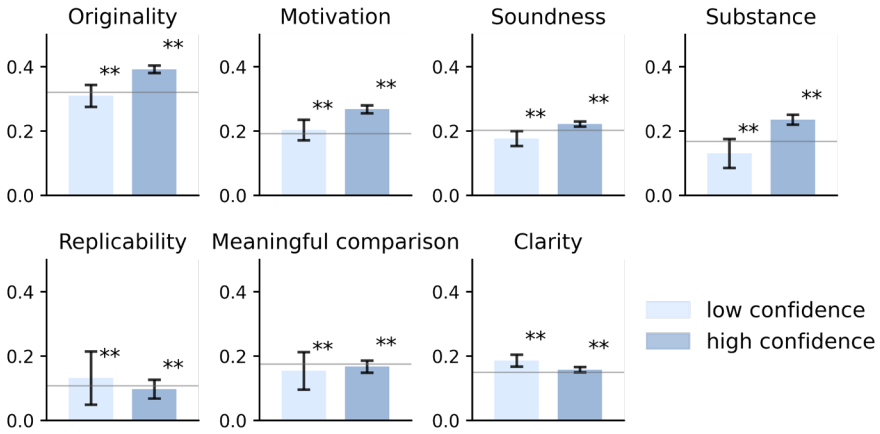


Figure 5. The impact of perceptions reflected in review content between high-confidence and low-confidence reviewers on their decisions (using the standardized median value, i.e., 5, as a cut-off, with the standardization process similar to *RD*). The middle line represents the coefficient for confidence at the median value of 5. The study found that, except for clarity and replicability, the perceptions of high-confidence reviewers are most closely associated with their decisions across various aspects. This reflects that decisions made by high-confidence reviewers are more interpretable and follow a logical progression from perception to decision-making.

Paper decision analysis (RQ3)

When deciding whether to accept a paper, the conference chairs take into account the reviewers' scores and feedback to make the final decision on acceptance or rejection. We further analyse how the chairs' decision is made to gain insights into the process of decision-making in peer review. X_p is the vector of control variables, including average *RD*, request number and the prestige of paper, where prestige refers to the average h-index of all authors, matched with S2AG (Version: May 21, 2024) (Wade, 2022). $PaperDecision_p$ represents the final decision for paper p. It is a binary variable representing whether the paper is accepted. The model is specified as follows:

$$PaperDecision_p = \beta_0 + \beta_1 RD_p + \beta PF_p + \beta' X_p \alpha_p + \lambda_p$$

We conducted a VIF test (Table S2) and used logistic regression for our analysis (Table 2). We uncovered an interesting distinction (Figure 6): from the perspective of the chairs, the most critical perceptual factors are not originality or motivation, but rather substance and soundness. This contrasts with the focus of reviewers. The findings suggest that chairs place greater emphasis on ensuring the overall validity in the field. Moreover, factors such as motivation, replicability, clarity, and meaningful comparison do not have a significant impact on paper decisions. This could be due to the supportive nature (as shown in Figure 3) of factors like motivation, which may lack strong differentiating power from the perspective of papers' decisions. It is shown in Figure 7 that in the high-*RD* group (RD below the median), positive judgments on validity-related dimensions, are more influential on papers' decisions.

Dependent Variable: Paper decision	Coefficient	SE	P value	Odds multiplier	Confidence interval
Motivation	0.078	0.052	0.132	1.08	[-0.024, 0.180]
Originality	0.110	0.047	0.020	1.12	[0.017, 0.203]
Soundness	0.184	0.034	0.000	1.20	[0.118, 0.250]
Substance	0.192	0.064	0.003	1.21	[0.066, 0.318]
Replicability	-0.172	0.123	0.160	0.84	[-0.413, 0.068]
Meaningful comparison	-0.067	0.079	0.397	0.94	[-0.222, 0.088]
Clarity	0.059	0.032	0.070	1.06	[-0.005, 0.122]
Reviewer Decision	2.810	0.056	0.000	16.60	[2.700, 2.919]
Request number	-0.005	0.009	0.603	1.00	[-0.023, 0.013]
Prestige	0.005	0.002	0.020	1.01	[0.001, 0.010]

Table 2. The regression results for paper decision analysis show that the pseudo-R-squared is 0.55, with a sample size of 17,620 observations. Since chairs' decisions are heavily influenced by the average scores provided by reviewers, we excluded papers rated below the 1st percentile of accepted submissions (a total of 5,464 papers, with a 97% probability of rejection) and those rated above the 99th percentile of rejected submissions (a total of 2,268 papers, with a 97% probability of acceptance).

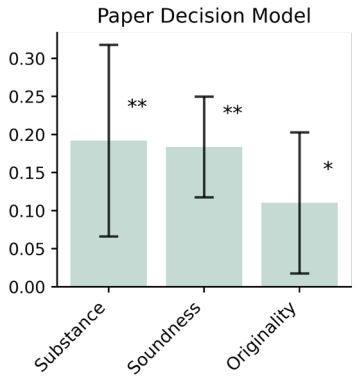


Figure 6. The coefficient comparison of paper decision analysis. We only present the main variables that are significant in the regression analysis. Among all the perceptual factors, substance and soundness have a significant positive correlation with the chair's decision, followed by originality. The effects of other main variables on the paper's decision are not significant. ** represents $p < 0.01$, and * represents $p < 0.05$.

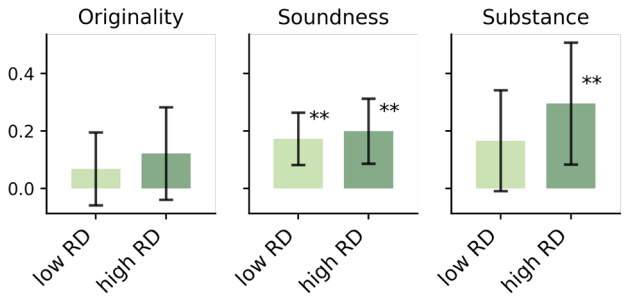


Figure 7. We compared the high and low *RD* groups (divided by the median, with extremely high and low scores excluded in the same way as in Table 2) regarding reviewers' perceptions of the papers. We found that decisions on papers in the high-*RD* group were more likely to be influenced by the reviewers' perceptions. High scores already indicate the overall superiority of the paper's quality. If reviewers' evaluations of the validity-related factors are positive, it indicates that the chairs' concerns have been addressed, thus increasing the likelihood that the chairs will accept the paper.

Conclusion and discussion

Due to limited review content and narrow analytical dimensions, there's little empirical research on the perception-to-decision process in peer review. Our findings show that originality is the most critical factor in reviewers' decisions, followed by motivation. Chairs rely on reviewers' decisions but prioritize assessments of soundness and substance, focusing on the research's overall validity. Reviewers' feedback on motivation is reassuring but has limited distinguishing ability for chairs. The Overall impact of perceived meaningful comparison, clarity, and replicability on decision-making in peer review is weak.

While many believe that novelty is always hampered in the publication process (Liang et al., 2023; Wang et al., 2017), our study empirically uncovers that reviewers place a substantial emphasis on their perception of originality and are more inclined to accept papers they deem novel. This highlights that originality is the most valued and recognized element by reviewers. However, the transition from perception to decision is complex and still largely unexplored. Specifically, reviewers' perspectives can be influenced by preconceptions or past experiences, and there is often a gap between perception and its expression. Additionally, cognitive limitations may obscure underlying social biases. The intricate mechanisms and overall effectiveness of this system demand ongoing exploration and refinement. Within this context, reviewers' cognition, perception, and decision-making are indispensable and critical components. While the peer review system benefits science, it also necessitates a continuous cycle of feedback to evolve and improve (Ward et al., 2015).

Limitation and future work

Due to challenges in accessing review content, our data has some constraints. We have also lightly addressed chairs' perceptions, which we will explore in more depth. The requests from reviewers also contain important information, and there is potential for further analysis.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (72104212).

About the authors

Yujie Zhang is a master student from the Department of Information Resources Management at Zhejiang University, China. They can be contacted at yj.zhang@zju.edu.cn.

Weikang Yuan is a doctoral student from the Department of Information Resources Management at Zhejiang University, China. They can be contacted at yuanwk@zju.edu.cn.

Zhuoren Jiang is an Assistant Professor the *corresponding author* of this paper from the Department of Information Resources Management at Zhejiang University, China. They can be contacted at jiangzhuoren@zju.edu.cn.

References

Alberts, B., Hanson, B., & Kelner, K. L. (2008). Reviewing Peer Review. Science, 321(5885), 15–15. https://doi.org/10.1126/science.1162115

Bailly, B. A. F. L. (2016). Learning from peer review. Nature Nanotechnology, 11 2, 204. https://doi.org/10.1038/nnano.2016.4

Balietti, S., Goldstone, R. L., & Helbing, D. (2016). Peer review and competition in the Art Exhibition Game. Proceedings of the National Academy of Sciences, 113(30), 8414–8419.

Benda, W. G., & Engels, T. C. (2011). The predictive validity of peer review: A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. International Journal of Forecasting, 27(1), 166–182.

Blockeel, C., Drakopoulos, P., Polyzos, N., Tournaye, H., & García-Velasco, J. (2017). Review the 'peer review'. Reproductive Biomedicine Online, 35 6, 747–749. https://doi.org/10.1016/j.rbmo.2017.08.017

Brosch, T., Scherer, K., Grandjean, D., & Sander, D. (2013). The impact of emotion on perception, attention, memory, and decision-making. Swiss Medical Weekly, 143(1920), w13786–w13786.

Chong, S. W., & Mason, S. (2021). Demystifying the process of scholarly peer-review: An autoethnographic investigation of feedback literacy of two award-winning peer reviewers. Humanities and Social Sciences Communications, 8(1), 1–11.

Curtin, P. A., Russial, J., & Tefertiller, A. C. (2018). Reviewers' Perceptions of the Peer Review Process in Journalism and Mass Communication. Journalism & Mass Communication Quarterly, 95, 278–299. https://doi.org/10.1177/1077699017736031

Cutler, S., Xia, Y., & Beddoes, K. (2022). Peering into the Black Box of Peer Review: From the Perspective of Editors. 9th Research in Engineering Education Symposium (REES 2021) and 32nd Australasian Association for Engineering Education Conference (REES AAEE 2021). https://doi.org/10.52202/066488-0095

Dror, I. E., & Fraser-Mackenzie, P. A. (2008). Cognitive biases in human perception, judgment, and decision making: Bridging theory and the real world. In Criminal investigative failures (pp. 79–94). Routledge.

García, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2015). Bias and effort in peer review. Journal of the Association for Information Science and Technology, 66(10), Article 10.

Hamilton, D., Fraser, H., Hoekstra, R., & Fidler, F. (2020). Journal policies and editors' opinions on peer review. eLife, 9. https://doi.org/10.7554/eLife.62529

Howard, L., & Wilkinson, G. (1998). Peer review and editorial decision-making. The British Journal of Psychiatry, 173(2), 110–113.

Hua, X., Nikolov, M., Badugu, N., & Wang, L. (2019). Argument Mining for Understanding Peer Reviews. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2131–2137.

Jefferson, T., Wager, E., & Davidoff, F. (2002). Measuring the quality of editorial peer review. Jama, 287(21), 2786–2790.

Lamont, M. (2009). How professors think: Inside the curious world of academic judgment. Harvard University Press.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2–17.

Liang, Z., Mao, J., & Li, G. (2023). Bias against scientific novelty: A prepublication perspective. Journal of the Association for Information Science and Technology, 74(1), 99–114.

Lindsay, P. H., & Norman, D. A. (2013). Human information processing: An introduction to psychology. Academic press.

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv Preprint arXiv:1907.11692.

Maddi, A., & Miotti, L. (2024). On the peer review reports: Does size matter? Scientometrics, 1–21.

Manchikanti, L., Kaye, A., Boswell, M., & Hirsch, J. (2015). Medical journal peer review: Process and bias. Pain Physician, 18 1, E1–E14.

Marietta, M., & Barker, D. (2019). The Psychology of Fact Perceptions II. One Nation, Two Realities. https://doi.org/10.1093/OSO/9780190677176.003.0006

Marsh, H. W., Bond, N. W., & Jayasinghe, U. W. (2007). Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. Australian Psychologist, 42(1), 33–38.

Matsui, A., Chen, E., Wang, Y., & Ferrara, E. (2021). The impact of peer review on the contribution potential of scientific papers. PeerJ, 9, e11999.

Mavrogenis, A. F., Quaile, A., & Scarlat, M. M. (2020). The good, the bad and the rude peer-review. In International Orthopaedics (Vol. 44, pp. 413–415). Springer.

Nobarany, S., & Booth, K. S. (2015). Use of politeness strategies in signed open peer review. Journal of the Association for Information Science and Technology, 66(5), 1048–1064.

Park, I.-U., Peacey, M. W., & Munafo, M. R. (2014). Modelling the effects of subjective and objective decision making in scientific peer review. Nature, 506(7486), 93–96.

Peters, D. P., & Ceci, S. J. (1982). Peer-review research: Objections and obligations. Behavioral and Brain Sciences, 5(2), 246–255.

Phillips, J. (2011). Expert bias in peer review. Current Medical Research and Opinion, 27, 2229–2233. https://doi.org/10.1185/03007995.2011.624090

Riley, B., & Jones, R. B. (2016). Peer review: Acknowledging its value and recognising the reviewers. The British Journal of General Practice : The Journal of the Royal College of General Practitioners, 66 653, 629–630. https://doi.org/10.3399/BJGP16X688285

Roediger, H. L. (1991). Is unreliability in peer review harmful? Behavioral and Brain Sciences, 14(1), 159–160.

Soergel, D., Saunders, A., & McCallum, A. (2013). Open Scholarship and Peer Review: A Time for Experimentation□.

Street, C. T., & Ward, K. (2019). Cognitive Bias in the Peer Review Process: Understanding a Source of Friction between Reviewers and Researchers. Data Base, 50, 52–70. https://doi.org/10.1145/3371041.3371046

Teplitskiy, M., Peng, H., Blasco, A., & Lakhani, K. R. (2022). Is novel research worth doing? Evidence from peer review at 49 journals. Proceedings of the National Academy of Sciences, 119(47), e2118046119.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. Proceedings of the National Academy of Sciences, 114(48), 12708–12713.

Travis, G. D. L., & Collins, H. M. (1991). New light on old boys: Cognitive and institutional particularism in the peer review system. Science, Technology, & Human Values, 16(3), 322–341.

Tvina, A., Spellecy, R., & Palatnik, A. (2019). Bias in the Peer Review Process: Can We Do Better? Obstetrics & Gynecology. https://doi.org/10.1097/AOG.0000000000003260

Wade, A. D. (2022). The semantic scholar academic graph (s2ag). Companion Proceedings of the Web Conference 2022, 739–739.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. Research Policy, 46(8), 1416–1436.

Ward, P., Graber, K. C., & Van Der Mars, H. (2015). Writing quality peer reviews of research manuscripts. Journal of Teaching in Physical Education, 34(4), 700–715.

Wenneras, C., & Wold, A. (2001). Nepotism and sexism in peer-review.

Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). Open peer review: Promoting transparency in open science. Scientometrics, 125(2), 1033–1051.

Yuan, W., Liu, P., & Neubig, G. (2022). Can we automate scientific reviewing? Journal of Artificial Intelligence Research, 75, 171–212.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Appendix

Supplementary materials

See supplementary materials.

Information Research - Vol. 30 No. iConf (2025)

How is AIGC shaping the world: an analysis of bibliometrics

Ying Guo, Fengchang Yu, Jianfei Lai, and Xiaoqun Yuan

DOI: https://doi.org/10.47989/ir30iConf47227

Abstract

Introduction. While artificial intelligence-generated content (AIGC) research is expanding, it largely focuses on specific domains, with limited analysis of its global impact across countries and disciplines. This paper seeks to fill that gap, offering insights into regional AIGC development for policymakers, investors, and researchers.

Data and Methods. Using bibliometric methods, this study analyses 11,638 AIGC-related publications from the period between ChatGPT's release on November 30, 2022, and July 31, 2024. A set of indicators is applied to assess the research performance globally, across continents, and in top-performing countries/regions.

Analysis and Results. Findings show global engagement in AIGC research, focusing on AI technological advancements, healthcare, education, and ethics. However, there is a notable lack of integration between AIGC and basic disciplines. AIGC research performance mirrors global economic development patterns, reflecting significant imbalances.

Conclusions. The study emphasizes the unequal global development of AIGC, calling for enhanced international cooperation, optimized resource allocation, and tailored strategies to close technological gaps.

Introduction

Artificial intelligence-generated content (AIGC) services are designed to support rapid and personalized multimodal content creation, along with a wide range of applications such as digital marketing, standard knowledge services, video game design, and filmmaking (Cao et al., 2023). By automating the generation of vast amounts of rich, personalized, high-quality content, AIGC has become a driving force for innovation, playing a key role in the next-generation internet paradigm, metaverse, and digital twin technologies. A study by PricewaterhouseCoopers estimates that AI could boost global GDP by 14%, or approximately $15.7 trillion, by 2030 (PwC, 2017). These figures underscore AIGC's transformative impact on economic growth and its role in accelerating technology adoption.

As an indispensable driving force in the transition to a new era of digital civilization, AIGC has injected fresh vitality into global technological and economic development, offering new opportunities in sectors like healthcare, education, and finance (Cooper, 2023; Dowling & Lucey, 2023; Khan et al., 2023; Lim et al., 2023; Thirunavukarasu et al., 2023). However, due to the inherent characteristics of AI models, coupled with economic, technological, and social factors, AIGC also introduces challenges such as ethics, privacy concerns, algorithmic bias, and regulatory oversight (Cotton et al., 2024; Dwivedi et al., 2023; Gao et al., 2023; Lund et al., 2023), Moreover, it exacerbates global imbalances in areas such as technology, economy, education, and resource distribution (Regmi, 2024; Wang et al., 2024). Therefore, it is crucial to explore how AIGC integrates across different regions and sectors, which can provide insights into the imbalances it creates and inform strategies for responding AIGC's role in shaping future societal development on a global scale.

While significant research exists on AIGC, much of it focuses on specific areas or perspectives such as banking (Beckmann & Hark, 2024; Chaouali et al., 2024; Kim & Song, 2024), libraries (Fruehauf et al., 2024; Lo, 2023; Oddone et al., 2024), and autonomous driving (Xu, Niyato, Chen, et al., 2023; Xu, Niyato, Zhang, et al., 2023; Zheng et al., 2023). Few studies, however, address the broader impact of AIGC across various countries and fields from a global perspective. This paper fills that gap by employing bibliometric methods to map the global landscape of AIGC-related research. It provides a macro-level analysis of the research performance of different continents and key countries in AIGC-related topics, offering insights into the development of AIGC across regions. These findings can inform regional policymaking, national strategic planning, economic investments, resource allocation, and international research collaboration.

Data and methods

In this paper, we firstly recognized the framework of the search formula by reviewing and backtracking the important literature, which not only contains AIGC, but also includes AIGC tools, technologies, tasks, and applications, etc.; thereafter, we supplemented the relevant keywords through the literature examination and expert consultation; and finally, through several rounds of adjustment, we obtained a search query containing about 100 search keywords, to maximize recall and ensure precision in the search for AIGC-related literature. Using the Web of Science Core Collection, a search was conducted on September 1, 2024, to retrieve relevant articles, reviews, and proceeding papers published between the release of ChatGPT on November 30, 2022, and July 31, 2024. This resulted in a dataset of 11,638 documents.

By comparing UN member states and their respective continents, this study identifies the country/region and its corresponding continent for each article. To better capture collaborative contributions from different countries/regions, we do not differentiate between the order of authors or the extent of their contributions. This means an article may be classified under multiple continents and countries/regions simultaneously. Based on the citation topics in InCites, each article in the dataset is categorized into unique macro-topics and micro-topics. Citation topics are

algorithmically derived citation clusters, developed by CWTS, Leiden. This system uses a three-level hierarchical classification, consisting of 10 macro-topics, 326 meso-topics, and 2,449 micro-topics. Macro- and meso-topics are manually labelled based on content, while micro-topics are algorithmically labelled by their most significant keywords. This classification method, with its clear and concise macro-topics and comprehensive, detailed micro-topics, is well-suited for macro- and micro-level analyses of AIGC literature across various continents and countries/regions.

In addition to the number of papers published, we have selected two normalized indicators as the main measures for the academic influence of each continent and country/region: impact relative to world (IRW) and category normalized citation impact (CNCI). To minimize the influence of extreme paper counts or citation frequencies, this study also uses journal normalized citation impact (JNCI), citation impact, documents in the top 1%, and top 10% as supplementary reference indicators (The meanings of each indicator can be found in https://incites.zendesk.com/hc/en-gb/sections/22667267182225-Indicators-and-Calculations).

Accordingly, the study will outline the global distribution of AIGC-related research, highlight the extent of AIGC's impact on different countries/regions, and offer decision-making support for stakeholders.

Results and discussion

Global overview

Figure 1 illustrates the global distribution of publications related to AIGC following the release of ChatGPT. The data indicate that, since then, an increasing number of nations have engaged in AIGC research. Specifically, over 80% of countries/regions in Asia and Europe have contributed to this body of work. African countries with relevant outputs in 2023 grew by 750%, with 24 African countries having published AIGC-related papers as of the data collection date. This trend demonstrates that AIGC research and applications are no longer concentrated in a few technology-developed regions but are spreading worldwide.

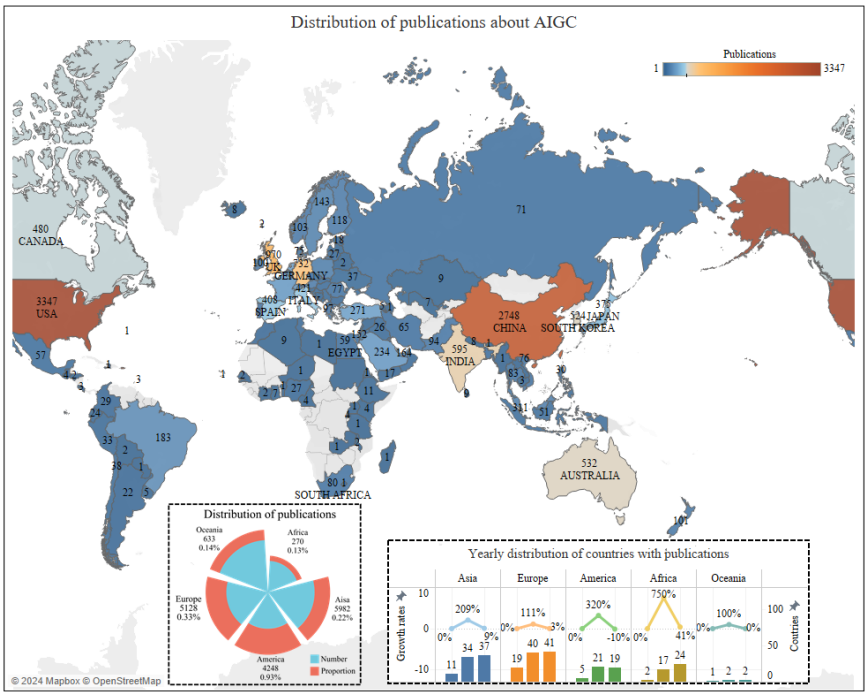


Figure 1. Distribution of publications about AIGC

The data suggest a regional imbalance in AIGC-related research. Asia and Europe have the highest number of publications, while the Americas have the highest percentage of AIGC-related papers, which is close to 1%. Notably, AIGC-related papers in Asia are mainly concentrated in China (46%), while in the Americas they are mainly concentrated in the United States (65%), reflecting the fact that AI technology has become the focal point of the global technology competition landscape. Although Oceania has fewer publications overall, its contribution from Australia and New Zealand is notable. In contrast, smaller nations in Oceania have yet to significantly engage in AIGC research, likely due to limited resources. Africa, with 266 papers from 24 countries, faces constraints in research resources and infrastructure, potentially slowing AIGC development. Nonetheless, many African nations are making efforts to explore AIGC's societal applications and build their technological capacities, even if only through an article.

Figure 2 illustrates the keywords co-occurrence network and the evolving trends of subject terms in global AIGC literature. Over time, researchers' focus has shifted from early terms like '*conversational agent*' to '*Chatbot*', and more recently to '*ChatGPT*', eventually broadening to '*generative* AI'. This progression reflects the advancement of human-computer interaction from basic conversational robots to sophisticated generative AI platforms, highlighting ongoing improvements in user experience, functional complexity, and application scenarios. Technologically, attention has transitioned from early concepts such as '*deep learning*', 'GAN (*generative adversarial networks*)', and 'CNN (*convolutional neural network*)' to 'NLP (*natural language Ppocessing*)' and '*transformer*' architectures, and now to 'LLM (*large language models*).' This shift demonstrates the evolution of AI from traditional deep learning methods to generative AI technologies centered around NLP, particularly the development of Transformer models and LLM.

High-frequency keywords indicate that since the release of ChatGPT, '*artificial intelligence*' has emerged as the dominant keyword in AIGC research. As a prominent application, 'ChatGPT' naturally draws significant attention and discussion within the field. Additionally, the increasing frequency of terms such as 'LLM' and '*generative* AI' in recent literature suggests a growing focus on the fundamental technologies and core architectures of generative AI. This shift highlights a transition in research from application-oriented tools to foundational technologies, signalling an expansion of AIGC studies from single model applications to a broader technical foundation and theoretical framework.

Red nodes around the high-frequency keywords highlight recent trends in AIGC research. On one hand, terms such as '*education*', '*higher education*', '*patient education*', and '*health*' indicate the expanding application of AIGC technology in education and medicine. This suggests that AIGC is extending beyond science and academic research into various societal domains, showing promise in addressing practical problems. On the other hand, the frequent co-occurrence of keywords like '*Prompt engineering*', '*task analysis*', '*performance*', '*accuracy*', '*quality*', and '*readability*' reflects a growing focus on the performance and precision of AIGC models. Researchers are exploring ways to enhance model output quality, improve content readability, and optimize task execution. This shift underscores the importance of performance optimization as AIGC technology advances from development to practical application. Additionally, the appearance of terms such as '*anxiety*', '*challenges*', and '*future*' indicates that while technological advancements are a focus, researchers are also considering potential challenges and social risks associated with AIGC. The rapid growth of generative AI raises issues related to ethics, security, privacy, and social anxiety, prompting scholars to contemplate future developments and address these concerns.

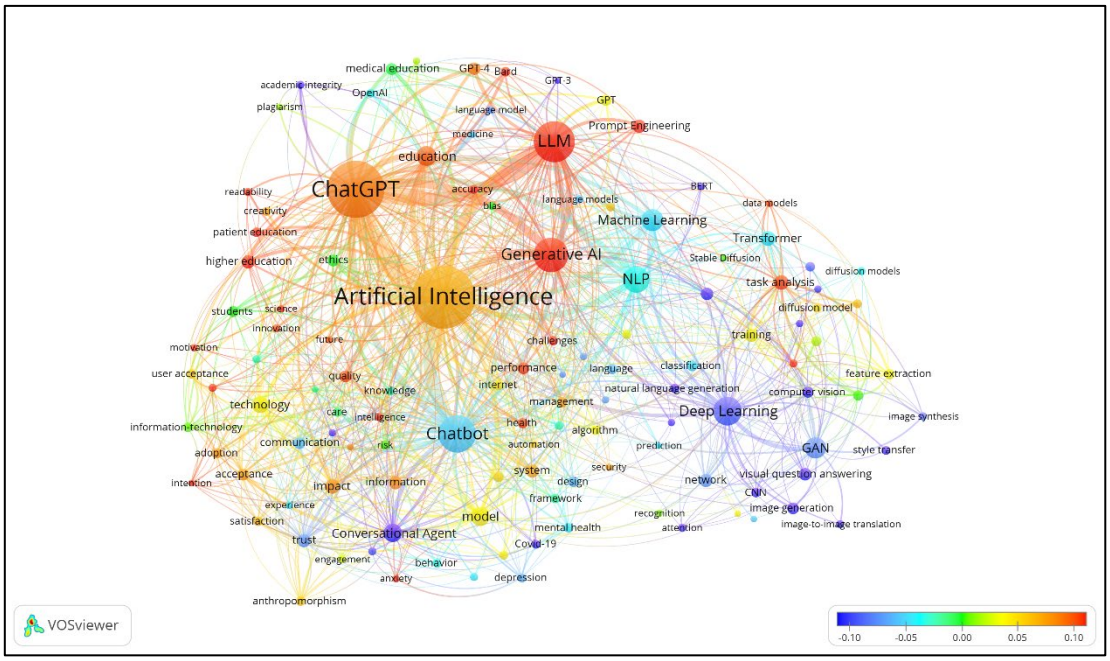


Figure 2. Keywords co-occurrence of publications about AIGC

The size of the circle indicates the frequency of the word; the colour distribution is the time evolution after standardization; the connecting line indicates the strength of the co-occurrence relationship.

Performance of continents in macro-topics

Figure 3 illustrates the distribution of AIGC-related research performance across continents and macro topics. The number of published papers indicates that AIGC technology has expanded beyond electrical engineering, electronics and computer science into various disciplines. However, its application in foundational fields such as mathematics and physics remains limited. This suggests that while AIGC technology has driven advancements in various disciplines, its potential in fundamental sciences could become a significant area of interest in the future, particularly for addressing complex mathematical modelling, theoretical validation, and physical simulation problems.

Examining the distribution of AIGC papers across disciplines by continent reveals that Asia leads in publication volume, followed by Europe and the Americas, with Oceania and Africa contributing fewer results. Asia's leading position may be attributed to its advancing scientific infrastructure, demographic advantages, and strong national policies. Europe and the Americas are notable for their technological innovations and focus on the application of AIGC in social sciences and clinical and life sciences, which reflects their emphasis on the social and ethical implications of AIGC. This focus is supported by robust science and technology policies, social security, and academic freedom in these regions. In Oceania, AIGC research has not yet made significant inroads into Mathematics, potentially due to a concentration of resources in specific application areas and relatively weaker basic theoretical research. Africa's AIGC research is emerging, with fewer publications across disciplines, and has yet to produce research in engineering and materials science, suggesting constraints in scientific resources and technological development in the region.

Africa and Oceania achieve the highest scores for AIGC research and the combination of electrical engineering, electronics and computer science in terms of IRW and CNCI. This outcome is contrary to initial expectations. Analysis of other indicators reveals that although Africa and Oceania publish fewer articles overall, the high-impact nature of their research results in a relatively high influence

around the world. A detailed review of the papers from these regions indicates that many are produced in collaboration with institutions in Asia, Europe, and the Americas. This suggests that, despite limited resources, enhancing international cooperation can significantly boost scientific research output and global impact. Meanwhile, the Americas excel across all indicators, demonstrating strong overall influence. Asia and Europe show similar scores in CNCI, but Asia has a higher number of highly cited papers and a greater proportion of papers in the Documents in the top 1% compared to Europe, highlighting Asia's superior performance in high-impact research.

In the fields of clinical and life sciences, the findings differ from the latter. Among Asia, Europe, and the Americas, where the volume of publications allows for comparison, the Americas lead in most impact indicators and normalized indicators, followed by Europe and Asia. Despite Oceania's relatively small number of publications, its proportion of high-impact papers remains notable, while Africa's performance across all indicators is comparatively lower.

In social sciences, Oceania's influence stands out distinctly. Despite having only 189 papers, Oceania has a proportion of Documents in the top 1% and top 10% of approximately 15.64% and 35.45%, respectively, significantly surpassing the other continents. Additionally, Oceania's CNCI is 4.28, the highest among the continents. Africa, with 71 publications in social sciences, has about 9.86% of its papers in the Documents in the top 1%, ranking just behind Oceania. Although Asia, Europe, and the Americas show close performance in social sciences, Europe exhibits a higher citation impact than Asia and the Americas, resulting in relatively higher IRW and CNCI for Europe.

Notably, among the 24 articles published in the field of mathematics in the Americas, the citation impact exceeds 10, and the proportions of Documents in the top 1% and top 10% are 25% and 37.5%, respectively, indicating a strong influence of the Americas in AIGC research within mathematics. The reference value of the measurements for other disciplines is limited due to the small number of papers.

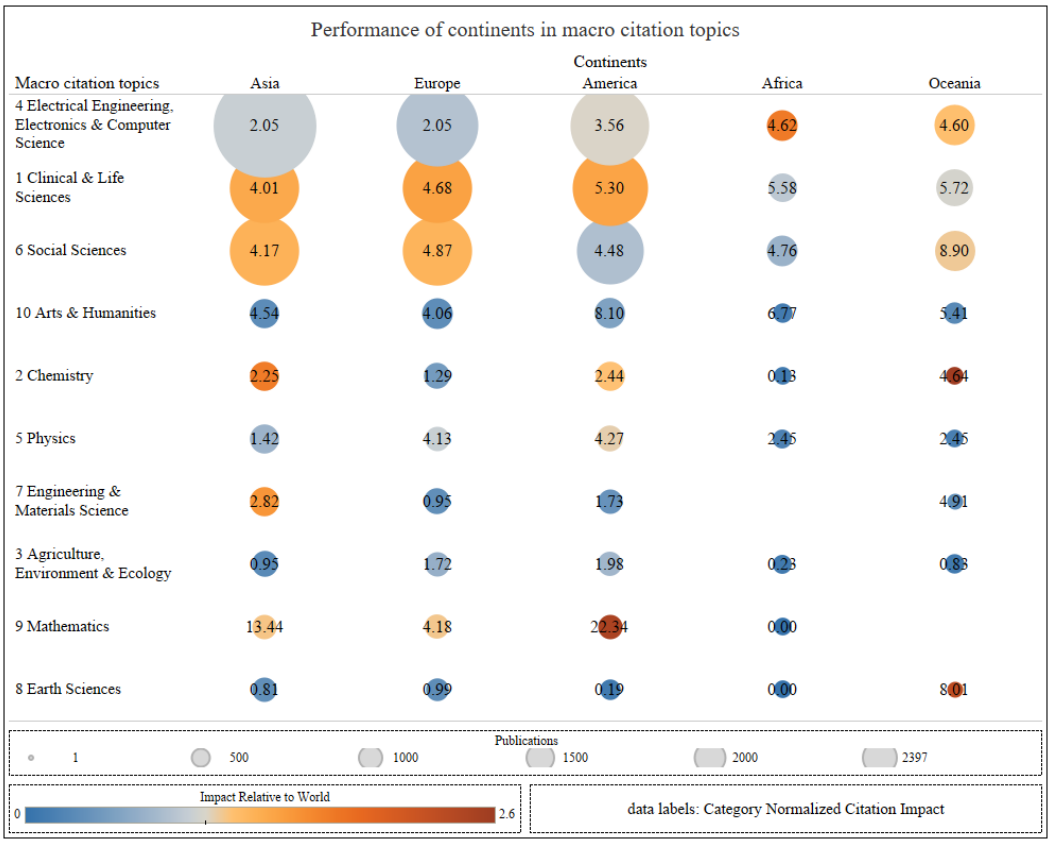


Figure 3. Performance of continents in macro citation topics

Performance of top countries in micro-topics

Based on the number of published papers, this article examines the top 10% of countries in each continent and identifies their top 5 micro-topics. The analysis reveals that research from countries in Asia, Europe, and the Americas predominantly centers on five micro-topics: deep learning, artificial intelligence in healthcare and medicine, human-robot interaction, natural language processing (NLP), and open data.

China leads in the number of publications within the field of deep learning. Additionally, NLP is a significant research topic in China, aligning with the country's rapid advancements in artificial intelligence applications and language models. However, in terms of citation impact, articles from China show slightly lower or similar levels compared to global and disciplinary averages, with a notable disparity when compared to the United States. This suggests that China still faces challenges in accumulating high-quality, high-impact research outcomes. This gap may be attributed to variability in paper quality or the fact that many highly cited papers originate from leading research institutions in other countries.

The United States exhibits exceptional citation influence, particularly in artificial intelligence in healthcare and medicine, as well as deep learning. Notably, in the deep learning field, the U.S. not only publishes a significant number of papers but also achieves substantial global impact. This underscores the United States' leading position in AI fundamental research and technology application, particularly in practical fields such as medicine and healthcare.

Unlike other countries in Asia, Europe, and the Americas, the UK does not feature open data among its top five research topics. Instead, it focuses on plagiarism, with notable citation performance in this area. This indicates that the UK's research on AI applications in academic ethics, anti-cheating

measures, and educational technology has garnered significant attention within the academic community.

Although African countries have fewer AIGC publications, they have increased their scientific research influence through international collaborations, particularly with Asia, Europe, and the Americas. Egypt's research aligns closely with that of European and American countries, while South Africa's focus areas differ. In addition to human-robot interaction and the technology acceptance model, South Africa's research also addresses specific issues such as health literacy, Sibutramine, and plagiarism.

Australia demonstrates strong citation influence in globally prominent topics such as deep learning, artificial intelligence in healthcare and medicine, and human-robot interaction. Additionally, Australia's research on plagiarism and Sibutramine is notable, reflecting its emphasis on academic integrity and health technology applications.

Although the number of papers published by different countries across various topics is relatively similar, citation influence varies significantly. For instance, India excels in human-computer interaction and open data. Italy and Spain have notable citation influence in human-computer interaction, with Spain's NLP papers particularly influential, though its deep learning papers are less impactful. These variations highlight differences in AIGC-related resource allocation, technological strengths, and research needs among countries, providing valuable insights into the global development landscape.

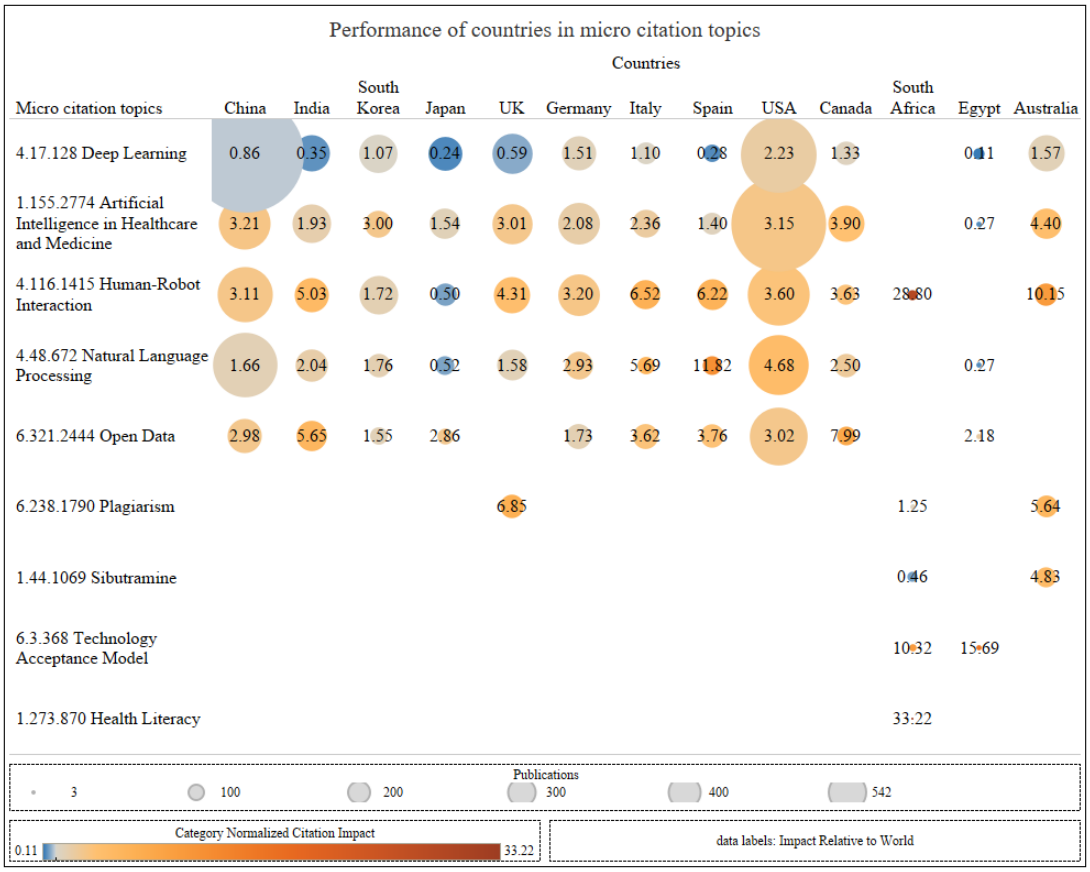


Figure 4. Performance of countries in micro citation topics

Conclusions

While AIGC has introduced disruptive technological changes across various industries and fields, it has also intensified global imbalances in technology, economy, education, and healthcare. Over the past two years, 123 countries/regions have published AIGC-related papers, primarily concentrated in three key fields: electrical engineering, electronics and computer science, clinical and life sciences, and social sciences. Asia, Europe, and the Americas exhibit notable advantages in research scale and international influence, with significant contributions from the United States and China in advancing AIGC technology. Oceania, though limited in research volume, demonstrates strong citation impact, while Africa generally lags behind.

As AIGC technology continues to evolve, adapting development strategies to align with its progress is crucial for decision-makers. Policymakers are expected to collaboratively develop global AIGC standards and ethical guidelines to foster technology sharing and address the digital divide. Enhancing multilateral cooperation through global platforms could help countries, particularly in Africa, access, and exchange research, potentially promoting more equitable global development. National strategies should consider local technological strengths; for instance, the U.S., Canada, and Australia might focus on advancing AIGC in medical fields and support international collaborations to improve technology levels. Investors should be attentive to emerging AIGC technologies like deep learning and NLP, as early investments might offer returns and assist in technology dissemination, especially in regions with less development. Researchers are advised to explore high-impact AIGC areas and integrate these with fundamental disciplines, while also strengthening international cooperation, particularly with African researchers, to support advancements in both technology and basic research.

Acknowledgements

This paper was supported by post-funding of National social science (research on empowering publishing knowledge services with digital intelligence), comprehensive construction technology of standard text resources and key elements (5400-202318585A-3-2-ZN), the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.72304215), and the Young Librarian Research Guidance Fund of Wuhan University Library (Grant No.2023-PY-02). In addition, the authors would like to acknowledge and thank Xia Liu, Ying Liu, Jia Lin, and Wenbo Tu of Wuhan University Library for their guidance and support in the earlier process of literature retrieval.

About the authors

Dr. Ying Guo is a Librarian in Wuhan University Library, Wuhan, China. She received her Ph.D. from Wuhan University, and her research interests are in the AIGC's impact on the future of libraries. She can be contacted at guoying49@whu.edu.cn.

Dr. Fengchang Yu is an Assistant Researcher in Information Management School, Wuhan University, Wuhan, China. He received his Ph.D. from Wuhan University, and his research interests are in the applications of AIGC in image recognition, etc. He can be contacted at yufc2002@whu.edu.cn.

Jianfei Lai is a Librarian in Wuhan University Library, Wuhan, China. Her research interests are in the AIGC's impact on the future of libraries. She can be contacted at jflai@whu.edu.cn.

Dr. Xiaoqun Yuan is an Associate Professor in Information Management School, Wuhan University, Wuhan, China. He received his Ph.D. from Huazhong University of Science and Technology, and his research interests are in the impact of AIGC on digital publishing. He is the corresponding author of this paper, and he can be contacted at yuan20030308@whu.edu.cn.

References

Beckmann, L., & Hark, P. F. (2024). ChatGPT and the banking business: Insights from the US stock market on potential implications for banks. Finance Research Letters, 63. https://doi.org/10.1016/j.frl.2024.105237

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-Generated Content (AIGC): A history of Generative AI from GAN to ChatGPT (arXiv:2303.04226). arXiv preprint.

Chaouali, W., Souiden, N., Aloui, N., Mouelhi, N. B. D., Woodside, A. G., & Abdelaziz, F. B. (2024). Roles of barriers and gender in explaining consumers' chatbot resistance in banking: A fuzzy approach. International Journal of Bank Marketing, 42 (7), 1867-1887. https://doi.org/10.1108/IJBM-03-2023-0153

Cooper, G. (2023). Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. Journal of Science Education and Technology, 32(3), 444–452. https://doi.org/10.1007/s10956-023-10039-y

Cotton, D., Cotton, P., & Shipway, J. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Dowling, M., & Lucey, B. (2023). ChatGPT for (Finance) research: The Bananarama Conjecture. Finance research letters, 53. https://doi.org/10.1016/j.frl.2023.103662

Dwivedi, Y., Kshetri, N., Hughes, L., Slade, E., Jeyaraj, A., Kar, A., Baabdullah, A., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M., Al-Busaidi, A., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). 'So, what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy. International Journal of Information Management, 71. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Fruehauf, E., Beman-Cavallaro, A., & Schmidt, L. (2024). Developing a foundation for the informational needs of generative AI users through the means of established interdisciplinary relationships. Journal of Academic Librarianship, 50 (3). https://doi.org/10.1016/j.acalib.2024.102876

Gao, C., Howard, F., Markov, N., Dyer, E., Ramesh, S., Luo, Y., & Pearson, A. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digital Medicine, 6(1). https://doi.org/10.1038/s41746-023-00819-6

Khan, R., Jawaid, M., Khan, A., & Sajjad, M. (2023). ChatGPT- Reshaping medical education and clinical management. Pakistan Journal of Medical Sciences, 39(2), 605–607. https://doi.org/10.12669/pjms.39.2.7653

Kim, D., & Song, H. (2024). Designing an age-friendly conversational AI agent for mobile banking: The effects of voice modality and lip movement. International Journal of Human-Computer studies, 187. https://doi.org/10.1016/j.ijhcs.2024.103262

Lim, W., Gunasekara, A., Pallant, J., Pallant, J., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarok or reformation? A paradoxical perspective from management educators. International journal of management education, 21(2). https://doi.org/10.1016/j.ijme.2023.100790

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. Journal of Academic Librarianship, 49(4). https://doi.org/10.1016/j.acalib.2023.102720

Lund, B., Wang, T., Mannuru, N., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. Journal of the association for information science and technology, 74(5), 570–581. https://doi.org/10.1002/asi.24750

Oddone, K., Garrison, K., & Gagen-Spriggs, K. (2024). Navigating Generative AI: The Teacher Librarian's Role in Cultivating Ethical and Critical Practices. Journal of the Australian Library and Information Association, 73(1), 3–26. https://doi.org/10.1080/24750158.2023.2289093

PwC. (2017, September 8). Sizing the prize: What's the real value of AI for your business and howcan you capitalise? PwC. https://www.pwc.com/gx/en/news-room/docs/report-pwc-ai-analysis-sizing-the-prize.pdf

Regmi, K. D. (2024). The rise of learning technology in an unequal world: Potentials and limitations in enhancing lifelong learning. International Review of Education, 70(3), 433–452. https://doi.org/10.1007/s11159-023-10058-2

Thirunavukarasu, A., Ting, D., Elangovan, K., Gutierrez, L., Tan, T., & Ting, D. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Wang, N., Wang, X., & Su, Y.-S. (2024). Critical analysis of the technological affordances, challenges, and future directions of Generative AI in education: A systematic review. ASIA Pacific Journal of Education, 44(1, S1), 139–155. https://doi.org/10.1080/02188791.2024.2305156

Xu, M., Niyato, D., Chen, J., Zhang, H., Kang, J., Xiong, Z., Mao, S., & Han, Z. (2023). Generative AI-Empowered Simulation for Autonomous Driving in Vehicular Mixed Reality Metaverses. IEEE Journal of Selected Topics in Signal Processing, 17(5), 1064–1079. https://doi.org/10.1109/JSTSP.2023.3293650

Xu, M., Niyato, D., Zhang, H., Kang, J., Xiong, Z., Mao, S., & Han, Z. (2023). Sparks of Generative Pretrained Transformers in Edge Intelligence for the Metaverse: Caching and Inference for Mobile Artificial Intelligence-Generated Content Services. IEEE Vehicular Technology Magazine, 18(4), 35–44. https://doi.org/10.1109/MVT.2023.3323757

Zheng, Z., Cheng, Y., Xin, Z., Yu, Z., & Zheng, B. (2023). Robust Perception Under Adverse Conditions for Autonomous Driving Based on Data Augmentation. IEEE Transactions on Intelligent Transportation Systems, 24(12), 13916–13929. https://doi.org/10.1109/TITS.2023.3297318

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Review articles as windows into Knowledge accumulation: the case of AI research

Siyuan Peng, Lei Hu, Jingrui Hou, Youqing Xia, and Ping Wang

DOI: https://doi.org/10.47989/ir30iConf47224

Abstract

Introduction. Review articles are essential in evolving scholarly information systems but have been underexplored in scientometrics. This paper aims to expand scientometric research on review articles, focusing on their role in understanding knowledge accumulation within specific domains.

Method. This study collected 4,315 review articles on artificial intelligence (AI). Using keyword frequency analysis and the Task-Technology Fit (TTF) model, the articles were classified into three categories: task-oriented, technology-oriented, and application-oriented.

Analysis. The temporal distribution of the review articles, the age distribution of their cited references, and the updating characteristics of references cited in review articles were analysed to provide preliminary insights into the evolution, dynamism, and updating patterns of knowledge in AI.

Results. The results show a marked increase in the publication frequency of review articles, especially over the past five years, with the application domain exhibiting the highest growth rate. Over half of the references cited in review articles across all domains are from the past five years. Additionally, older references are cited more frequently in newer reviews than more recent ones.

Conclusion. This study can be seen as an expansion of scientometric research based on review articles and highlights several intriguing research questions for exploration within this field.

Introduction

Recent decades have witnessed exponential growth in the volume of new scientific knowledge (Park et al., 2023). This incredible rate of scientific productivity leads to information overload, making it hard to keep up with state-of-the-art and assess the collective evidence in a particular research area (Snyder, 2019). By integrating diverse knowledge streams in a particular area, the genre of the review article could provide a solution. Review articles are critical evaluations of prior studies that have already been published (Palmatier et al., 2018). They are imperative for consolidating knowledge, synthesizing empirical findings, developing novel theories, comprehending the depth and breadth of research phenomena, and identifying unexplored areas for future investigation (Farrukh & Sajjad, 2023). Consequently, review articles have gradually gained recognition as important objects of scholarly research in the field of scientometrics (Blümel & Schniedermann, 2020). Most studies have focused on the structural features of review articles, such as the number of co-authors, pages, and patterns of institutional collaboration (Barrios et al., 2013). Other studies have explored the impact of review articles, including their influence on journal rankings and metrics (Ketcham & Crawford, 2007; Fassin, 2021), subsequent citations of cited papers (Lachance et al., 2014), citation-based analysis (Ho et al., 2017), and the measurement of individual research output (Miranda & Garcia-Carpintero, 2018). However, scientometric studies on review articles have generally been less often executed for more exploratory purposes (Blümel & Schniedermann, 2020), particularly when it comes to understanding the phenomenon of knowledge accumulation.

Review articles hold significant potential as a genre for exploring knowledge accumulation, defined as the collective body of knowledge gathered over time (Chandra & Dong, 2018). This potential arises from the fact that scientific knowledge in a particular field is often published in research articles, while review articles serve as a means of integrating and synthesizing these research articles (Praus, 2023; Snyder, 2019). Consequently, this paper aims to broaden the research scope regarding review articles in scientometrics by examining their role in elucidating the phenomenon of knowledge accumulation. We identify three potential research directions that warrant further exploration: The first is ***temporal distribution of review articles***. As review articles synthesize and present the current state of knowledge, their temporal distribution can serve as an indicator of the field's evolutionary pace and the periods of significant theoretical or empirical advancements (Patriotta, 2020). Analyzing the frequency and timing of review articles could reveal patterns of knowledge maturation and periods of heightened scholarly interest or innovation. The second is ***age distribution of references cited in review articles***, which investigates the age profile of references cited in review articles. Since the age of cited references is indicative of the rate of growth of scientific knowledge (Stacey, 2021), this paper proposes that the age distribution of references cited in review articles can, to some extent, reveal the knowledge accumulation and dynamism in the field. Finally, the third is ***updating characteristics of references cited in review articles***, which explores the patterns of reference updating in review articles. Review articles are regarded as an expansion and update of a previously published article on this topic (Torraco, 2016). Based on this viewpoint, this paper reckons that examining the updating patterns of references in review articles can further help reveal the knowledge accumulation speed and pattern.

As a pioneering study, this study provides some preliminary phenomena and trend disclosures within the three potential research directions to enhance our understanding of the accumulation of domain knowledge, and to expand research on review articles in scientometrics. Moreover, given the accelerating pace of knowledge accumulation within the field of artificial intelligence (AI) research and the growing number of review articles providing a comprehensive overview of this field (Leeming, 2021), we have chosen the review articles in the AI domain to conduct our analysis.

Data and preliminary investigation

Data preparation

Our research focused on the '*computer science, artificial intelligence*' category from the Web of Science (WoS), considering literature published up to the end of 2023. We initially obtained 5,991 results based on the review document type assigned by WoS. However, due to potential misclassification of review articles in WoS (Blümel & Schniedermann, 2020), this study underwent automatic preprocessing and manual review to exclude incomplete or non-review articles. As a result, we compiled a final set of 4,315 AI review articles.

To identify major research themes within this body of literature, we performed a keyword frequency analysis, allowing us to systematically uncover prevailing topics and trends discussed in the AI review articles. These findings were then structured with the task-technology fit (TTF) model (Goodhue & Thompson, 1995), which categorizes research topics based on their focus on either tasks, technology, or applications (i.e., the utilization of TTF). Accordingly, we classified the identified topics into three broad categories:

- **Task-oriented reviews:** these reviews emphasize AI tasks that are central to various domains. Examples include tasks such as '*classification*', '*pattern recognition*', and '*clustering*'. These reviews often concentrate on algorithmic methodologies and performance evaluation specific to task-related challenges.
- **Technology-oriented reviews:** in this category, the reviews focus on specific AI technologies, such as '*feature extraction*', '*convolutional neural networks (CNNs)*', and '*reinforcement learning*'. These reviews aim to provide a comprehensive understanding of cutting-edge AI methodologies and their theoretical underpinnings.
- **Application-oriented reviews:** this category explores the application of AI techniques to real-world challenges, often across diverse sectors, such as '*COVID-19 diagnostics*', '*medical imaging*' and '*autonomous systems*.' These reviews assess how AI technology, when aligned with the TTF model, can drive innovation, and provide solutions to complex, domain-specific problems.

Temporal distribution characteristics of review articles

This study mapped the temporal trends of AI review articles across three domains (see Figure 1), observing a general increase in publication frequency over time, especially notable since 2019. This aligns with the rapid growth of scientific literature noted by Bornmann et al. (2021). Then, we introduced '*density*' as a measure for the frequency of articles per year, identifying the '*COVID-19*' topic within the application domain as having the highest density (19.5) despite its brief existence. Other topics like '*internet of things (IoT)*', '*generative adversarial networks (GAN)*', '*blockchain,*' and '*federated learning*' have also shown high densities and an upward knowledge accumulation trend. Moreover, traditional task domain topics like '*classification*' and '*object detection*' demonstrate high densities with stable development.

As shown in Table 1, the comparative analysis of the past five years and five years prior revealed that the task domain typically showed higher article densities than the technology domain, which outpaced the application domain. The last five years witnessed significant density increases across all domains. Annual growth rates across these domains have converged and stabilized post-2019, with the application domain experiencing a marked growth rate increase after 2019, while the task and technology domains maintained consistent growth rates before and after 2019, stabilizing thereafter.

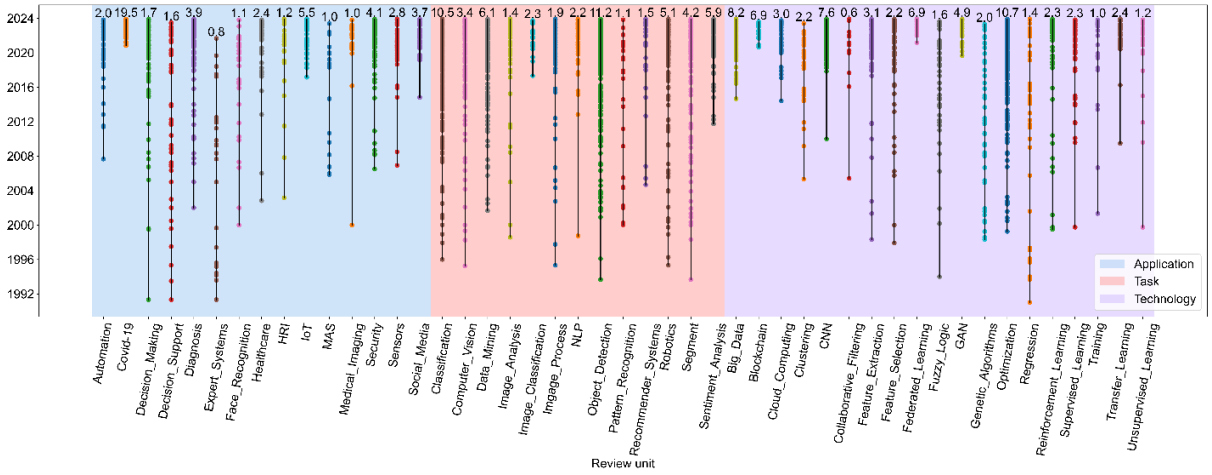


Figure 1. The temporal distribution of review article publications. The vertical span of each line represents the time span of publication for each review unit. Each point on the line segment represents a review paper emerging at a particular time point.

Review domain	Density (Overall)	Density (After 2019)	Density (Before 2019)	AAGR (After 2019)	AAGR (Before 2019)
Application	19.67	97.52	6.18	0.30±0.04	0.17±0.15
Task	47.21	201.85	17.27	0.27±0.03	0.27±0.30
Technology	32.27	156.68	10.46	0.29±0.03	0.25±0.24

Table 1. Density and average annual growth rate (AAGR) in different review domains

Age distribution characteristics of references cited in review articles

This study delved into the age distribution of references cited in AI review articles across three domains, defining '*reference lag years*' as the time difference between the publication year of the review and its cited literature. Figure 2 revealed a predominant lag of 2 to 3 years in task and technology domains, with the application domain occasionally extending to 4 years. Post-peak, all domains showed a rapid decline in citation numbers with increasing lag time.

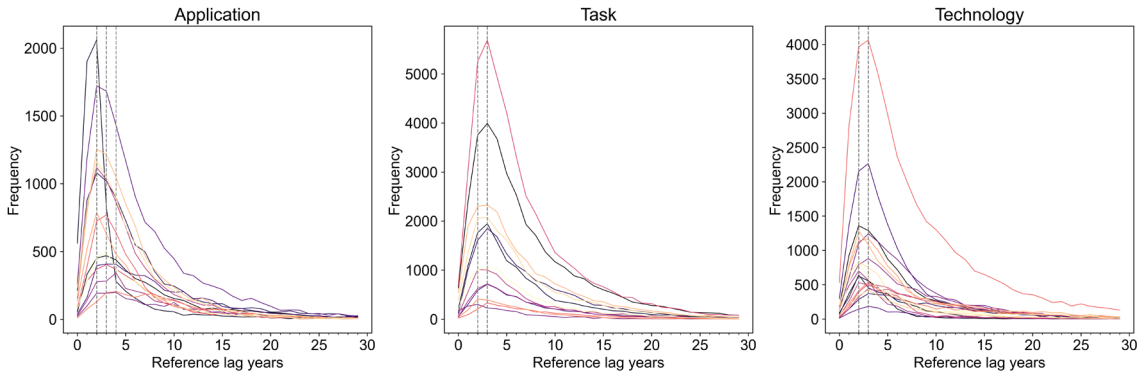


Figure 2. The distribution pattern of reference lag years in different review domains

A closer examination of reference lag years within a five-year span showed distinct patterns as shown in Table 2, the application domain's citations peak in the second year (0.136), the task domain in the third year (0.123), and the technology domain equally in the second and third years (0.123). The application domain demonstrated notably higher citation frequencies in the first two years compared to task and technology, suggesting quicker incorporation of recent literature.

Across all three domains, citations with a lag of up to five years constituted over half of the total, with the highest proportion in the application domain.

Review domain	Year 1	Year 2	Year 3	Year 4	Year 5	Sum
Application	0.125	0.136	0.124	0.106	0.084	0.575
Task	0.086	0.114	0.123	0.108	0.090	0.521
Technology	0.095	0.123	0.123	0.105	0.085	0.531

Table 2. Proportion of citations within recent five years for each review domain

In addition to analysing the relative reference lag years, this paper also examined the overall citation time distribution of review articles across different years. As shown in Figure 3, recent review articles span longer historical periods than older reviews. Specifically, literature time spans starting years shifted significantly backward from 1971, 1962, and 1958 in the year 2000 to before 1950 by 2022 and 2023 for the application, task, and technology domains respectively. Early 2000s and recent reviews showed more concentrated interquartile ranges, with wider ranges and noticeable fluctuations in other periods.

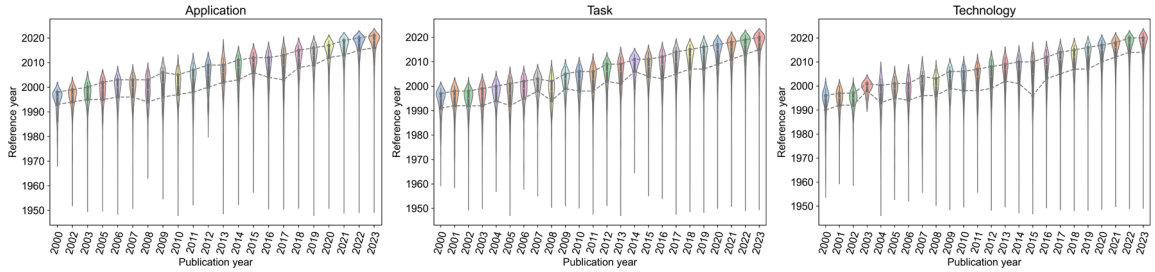


Figure 3. The distribution of publication years of references in review articles for different publication years

Updating characteristics of references cited in review articles

In examining the dynamics of references cited in AI review articles, we define references as '*knowledge units*' with temporal attributes, together forming a '*knowledge set*' over time. Our assessment of knowledge accumulation for each year's new references is threefold: (1) **accumulated references**: the cumulative total of references cited in all review articles up to a given year, indicating the knowledge set's size to that date. (2) **New references:** references cited for the first time in that year, expanding the existing knowledge set. (3) Types of new references: distinguished as either **recent** (published within the past year) or **previous** (published more than a year ago).

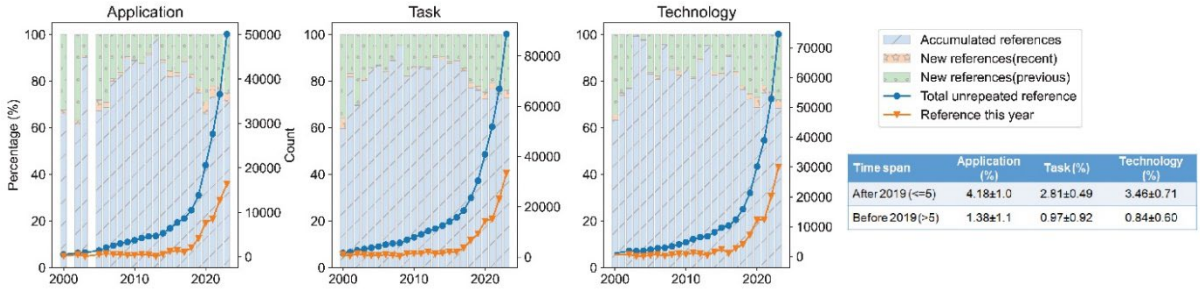


Figure 4. The annual changes in accumulated references and new references and proportion of new references (recent) in each review domain

As shown in Figure 4, the ratio of accumulated to new references indicates a predominance of accumulated references annually. Within new references, those published over a year ago consistently outnumber recent publications. Nevertheless, the share of recent references within

new citations has been rising lately. Quantitative analysis shows that the proportion of recent references in new citations post-2019 has increased across all domains.

Discussion and conclusion

Key findings

This study investigates the dynamics of knowledge accumulation in scientific fields by analysing review articles across multiple dimensions. Using AI research as a case study, we uncover patterns that are likely applicable to other evolving domains. Our analysis reveals several noteworthy findings and insights. Firstly, with regard to the temporal distribution characteristics, there has been a significant surge in their publication frequency, particularly in the past five years. This surge indicates that the pace of knowledge accumulation in AI is rapidly accelerating, aligning with the explosive growth of AI publications (Leeming, 2021). The domain of application has experienced the highest growth rate in terms of review articles, which can be attributed to the emergence of AI as a powerful tool for addressing complex interdisciplinary problems across various research fields (Zhuang et al., 2020). Secondly, in terms of the age distribution characteristics of references, more than half of the references cited in review articles across all domains fall within a five-year timeframe, with the application domain exhibiting the highest proportion. This trend signifies a notable acceleration in knowledge accumulation across the domains, especially in the application domain over the past five years, which aligns with the observed temporal distribution patterns. Furthermore, the analysis demonstrates that recent review articles encompass a broader timeframe of references. This finding is consistent with the nature of scientific knowledge accumulation, which involves the growth of the available stock of knowledge over time (Petrovich, 2018). Thirdly, in relation to the updating patterns of cited references, the number of older references cited in newer reviews consistently exceeds that of more recent references. This observation confirms the existence of a citation lag and suggests a pattern of delayed citation across review articles. Moreover, the increasing proportion of recent references among new citations in recent years indicates a heightened pace of knowledge accumulation in the three domains under examination.

Despite these insightful findings, the limitations should be acknowledged. Our analysis focuses on the AI research field. The findings may not generalize to other disciplines. Future research can explore different fields to help further clarify the role of review articles in understanding the knowledge accumulation phenomenon.

Conclusion and future work

Overall, this study presents some findings on knowledge accumulation in the field of AI based on a quantitative analysis of review articles. Based on these findings, this paper identifies several intriguing research questions that can be explored in the context of review-based scientometric studies.

The convergent growth of review articles and research articles. This question arises from the intuitive notion that review articles serve as summaries and evaluations of research articles (Praus, 2023) and the observed increase in the publication frequency of review articles, alongside a similar trend in research articles (Leeming, 2021). A promising avenue for further investigation is to delineate the quantitative relationship between the expansion of review articles and the overall growth of research articles.

When will a new review article emerge in a particular knowledge domain? Prior research suggests that the presence of review articles reflects the maturity of scientific fields (Bastide, 1989) and may manifest even in the early stages of their formation (Blümel, 2016), a notion supported by our observations (see Figure 1). Therefore, examining the predictors of when a new review paper will surface, based on the existing structure of research articles, could yield valuable insights.

Rethinking knowledge accumulation via both the citation pattern and semantic evolution of review articles. Leveraging review articles to explore the evolution of domain knowledge can provide valuable insights into the dynamic changes in knowledge development. While current analyses rely primarily on references and citation lags, future research should incorporate textual data and semantic evolution for deeper insights into patterns of knowledge accumulation.

In conclusion, this study not only enhances our understanding of knowledge accumulation in AI but also raises broader questions that are central to scientometrics and information studies. By advancing methods for tracking and analyzing review articles, this work establishes a foundation for future research into how scientific knowledge evolves across diverse disciplines. It suggests that review article analysis can be further developed to capture the dynamics of research development more effectively. Moreover, expanding the scope to include additional disciplines will enable future studies to refine existing models of knowledge diffusion and scholarly communication, offering deeper insights into the processes of knowledge creation and dissemination across scientific domains.

Acknowledgements

This work was supported by the National Social Science Fund of China (Grants No. 24&ZD186).

About the authors

Siyuan Peng is a Doctoral student at the School of Information Management, Wuhan University. Her research interests lie in scientometrics and knowledge mining. She can be contacted at siyuanpeng@whu.edu.cn.

Lei Hu is a master's student at the School of Microelectronics, Hubei University. His current research interests include data mining, deep learning, intelligent document understanding, and large language models. He can be contacted at hulei org@126.com.

Jingrui Hou is the corresponding author of this paper. He is a Postdoctoral researcher at the School of Information Management, Wuhan University, with research interests in natural language processing and knowledge evolution analysis. He can be contacted at jhou.research@outlook.com.

Youqing Xia received his master's degree from The University of Queensland, Australia. His research interests include the interdisciplinary application of archival studies and the processing of complex data. He can be contacted at xiayq519@gmail.com.

Ping Wang is a Professor at the School of Information Management, Wuhan University. He received his Ph.D. from Wuhan University, and his research interests include human-computer interaction and natural language processing. He can be contacted at wangping@whu.edu.cn.

References

Barrios, M., Guilera, G., & Gómez-Benito, J. (2013). Impact and structural features of meta-analytical studies, standard articles, and reviews in psychology: Similarities and differences. Journal of Informetrics, 7(2), 478-486. https://doi.org/10.1016/j.joi.2013.01.012

Bastide, F., Courtial, J., & Callon, M. (1989). The use of review articles in the analysis of a research area. Scientometrics, 15(5-6), 535-562. https://doi.org/10.1007/bf02017070

Blümel, C., & Schniedermann, A. (2020). Studying review articles in scientometrics and beyond: a research agenda. Scientometrics, 124(1), 711-728. https://doi.org/10.1007/s11192-020-03431-7

Blümel, C. (2016). Enrolling the toggle switch: Visionary claims and the capability of modeling objects in the disciplinary formation of synthetic biology. NanoEthics, 10, 269-287. https://doi.org/10.1007/s11569-016-0276-z

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications, 8(1), 1-15. https://doi.org/10.1057/s41599-021-00903-w

Chandra, P., & Dong, A. (2018). The relation between knowledge accumulation and technical value in interdisciplinary technologies. Technological Forecasting and Social Change, 128, 235-244. https://doi.org/10.1016/j.techfore.2017.12.006

Farrukh, A., & Sajjad, A. (2023). A critical review of literature review methodologies. Advancing Methodologies of Conducting Literature Review in Management Domain, 2, 103-123. https://doi.org/10.1108/S2754-586520230000002006

Fassin, Y. (2021). The impact of review articles in management and economics journal rankings and metrics. Scientometrics, 126(12), 9623-9632. https://doi.org/10.1007/s11192-021-04160-1

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. MIS quarterly, 213-236. https://doi.org/10.2307/249689

Ho, M. H. C., Liu, J. S., & Chang, K. C. T. (2017). To include or not: the role of review papers in citation-based analysis. Scientometrics, 110, 65-76. https://doi.org/10.1007/s11192-016-2158-0

Ketcham, C.M., & Crawford, J.M. (2007). The impact of review articles. Laboratory Investigation, 87, 1174-1185. https://doi.org/10.1038/labinvest.3700688

Lachance, C., Poirier, S., & Larivière, V. (2014). The kiss of death? The effect of being cited in a review on subsequent citations. Journal of the Association for Information Science and Technology, 65(7), 1501-1505. https://doi.org/10.1002/asi.23166

Miranda, R., & Garcia-Carpintero, E. (2018). Overcitation and overrepresentation of review papers in the most cited papers. Journal of Informetrics, 12(4), 1015-1030. https://doi.org/10.1016/j.joi.2018.08.006

Leeming, J. (2021). AI and its growing foothold in the natural sciences. Nature, 598, S5. *https://doi.org/10.1038/d41586-021-02762-6*

Palmatier, R. W., Houston, M. B., & Hulland, J. (2018). Review articles: Purpose, process, and structure. Journal of the Academy of Marketing Science, 46, 1-5. https://doi.org/10.1007/s11747-017-0563-4

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. Nature, 613(7942), 138-144. https://doi.org/10.1038/s41586-022-05543-x

Patriotta, G. (2020). Writing impactful review articles. Journal of Management Studies, 57(6), 1272-1276. https://doi.org/10.1111/joms.12608

Petrovich, E. (2018). Accumulation of knowledge in para-scientific areas: The case of analytic philosophy. Scientometrics, 116(2), 1123-1151. https://doi.org/10.1007/s11192-018-2796-5

Praus, P. (2023). Empirical relationship between the number of review and research articles. Scientometrics, 128(4), 2201-2209. https://doi.org/10.1007/s11192-023-04654-0

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of business research, 104, 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stacey, A. G. (2021). Ages of cited references and growth of scientific knowledge: an explication of the gamma distribution in business and management disciplines. Scientometrics, 126(1), 619-640. https://doi.org/10.1007/s11192-020-03761-6

Torraco, R. J. (2016). Writing integrative literature reviews: Using the past and present to explore the future. Human resource development review, 15(4), 404-428. https://doi.org/10.1177/1534484316671606

Zhuang, Y., Cai, M., Li, X., Luo, X., Yang, Q., & Wu, F. (2020). The next breakthroughs of artificial intelligence: The interdisciplinary nature of AI. Engineering, 6(3), 245-247. https://doi.org/10.1016/j.eng.2020.01.009

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Exploring the patterns and influencing factors of emerging technology impact: a case study of digital medical technology

Man Jiang, Siluo Yang, and *Shuang Tao*

DOI: https://doi.org/10.47989/ir30iConf47221

Abstract

Introduction. Impact, as an intrinsic attribute of emerging technologies, plays a crucial role in their identification and understanding. Traditional citation metrics, while reflective of impact magnitude, fail to explain its nature.

Method. Addressing this deficiency, this study categorizes the impact of emerging technologies into internal and external field, as well as impacts on basic versus applied development. With a focus on digital medical technology, this research investigates these four distinct impact patterns and their determinants. Using ordinary least squares (OLS) regression analysis on patent records from the Derwent innovations index (DII), the study examines how temporal novelty, innovation novelty, growth, and uncertainty influence both the magnitude and patterns of technological impact.

Analysis. Analysis shows a U-shaped relationship between temporal novelty and impact, indicating that older technologies gain influence through cumulative effects, while emerging technologies attract attention early on due to their pronounced novelty. Additionally, innovation novelty positively correlates with impact, underscoring its critical role in the effectiveness of emerging technologies. In contrast, growth rates and uncertainty demonstrate inconsistent effects.

Conclusions. These insights offer valuable guidance for policymakers, investors, and R&D managers in strategically advancing the development and deployment of emerging technologies.

Introduction

In our rapidly evolving digital era, emerging technologies such as digital medical technology are not only reshaping the healthcare ecosystems but also the broader socio-economic structures in which they operate. This transformation is particularly pronounced in the age of '*AI-algorithmic*' control, where the advancement of technologies like text mining and natural language processing (NLP) enhances our ability to identify and assess the impacts of emerging technologies more precisely and deeply. Understanding the impact patterns and influencing factors of these technologies becomes crucial, as they are often pivotal in driving innovation and boosting productivity (Rotolo et al., 2015). Given their potential to lead new industrial revolutions, reshape existing industry structures, optimize resource allocation, and propel economic and social advancement, it is essential, with the aid of text mining techniques, to conduct a comprehensive study of their impacts.

Existing studies predominantly focus on two areas: evaluating the impact of technologies and identifying the factors that drive these impacts. Methods such as bibliometrics and text mining have been extensively used to develop indices and models to assess technological influence (Gao et al., 2021). Citation metrics, widely regarded as appropriate indicators, suggest that a higher number of citations implies greater technological impact (Lee et al., 2012). However, these indicators merely quantify the extent of influence and fail to elucidate the nature of the impacts themselves. On the other hand, studies on factors impacting technology often focus on its role in the emergence of new industries or upgrades within existing ones, identifying key elements that yield significant technological impacts, typically using patent citations as metrics targeted towards enterprises or industries (Messeni Petruzzelli et al., 2015). While these studies offer valuable insights, they frequently overlook the nature of technological impacts and the diverse factors influencing various impact types. This oversight is critical as it limits our comprehension of how attributes of emerging technologies can lead to distinct patterns of influence, pivotal for shaping and transforming technological evolution paths.

The impacts of emerging technologies manifest in one or several domains and produce extensive cross-domain effects, influencing the entire system of technological innovation. Different types or patterns of influence have varied effects on technological innovation, necessitating targeted technological strategies by enterprises. This requires an analysis based on subdivided technological impact patterns to predict technological developments and mitigate R&D investment risks. However, current research has not systematically differentiated these technological impact patterns nor explored the factors influencing different patterns. Addressing these gaps, this study seeks to explore and resolve the following questions:

> Q1. How can we categorize technological impact patterns to consider the nature of their influences?
>
> Q2. Do attributes of emerging technologies affect different impact patterns in distinct ways?
>
> Q3. Which attributes significantly influence specific impact patterns, and what are the underlying mechanisms?

By exploring these questions, this study aims to enhance the impact of emerging technologies through innovative means, thus promoting their development and providing robust support for a deeper analysis of the intrinsic mechanisms influencing emerging technologies.

Method

Building upon prior studies, this research utilizes data from 15,116 patent records in the digital medical technology sector derived from the Derwent innovations index (DII). The abstracts of these patents were processed using natural language processing (NLP) to extract technological

terminologies, detailed in Jiang et al. (2024). Based on these findings, 1016 emerging technology terms within the digital medical technology field were selected using threshold indices (https://github.com/jiangman29/Paper-Data_JOI_ETs-Identification.git). These terms, characterized by their ET-scores and attributes of novelty (including temporal and value novelty), growth, uncertainty, and impact, serve as the sample for this study. Given the innovative and interdisciplinary nature of digital medical technology—which intersects medical, AI, and telecommunication fields—the impact patterns of this technology represent crucial strategic information for technology management. Technological innovations in this context play a key role in enterprise performance and innovation management.

The aim of this study is to explore the patterns of technological impact of emerging technologies and to analyse how attributes of these technologies influence their impact under different impact patterns. To achieve this, we first categorize the impacts of emerging technologies based on their nature into impacts within the technological field and impacts beyond the technological field, as well as impacts on basic and applied research. We then employ ordinary least squares (OLS) regression analysis to examine how the attributes of novelty, growth, and uncertainty influence the impacts across these four patterns. The overall framework is depicted in Figure 1. The subsequent section will detail the procedures of each step.

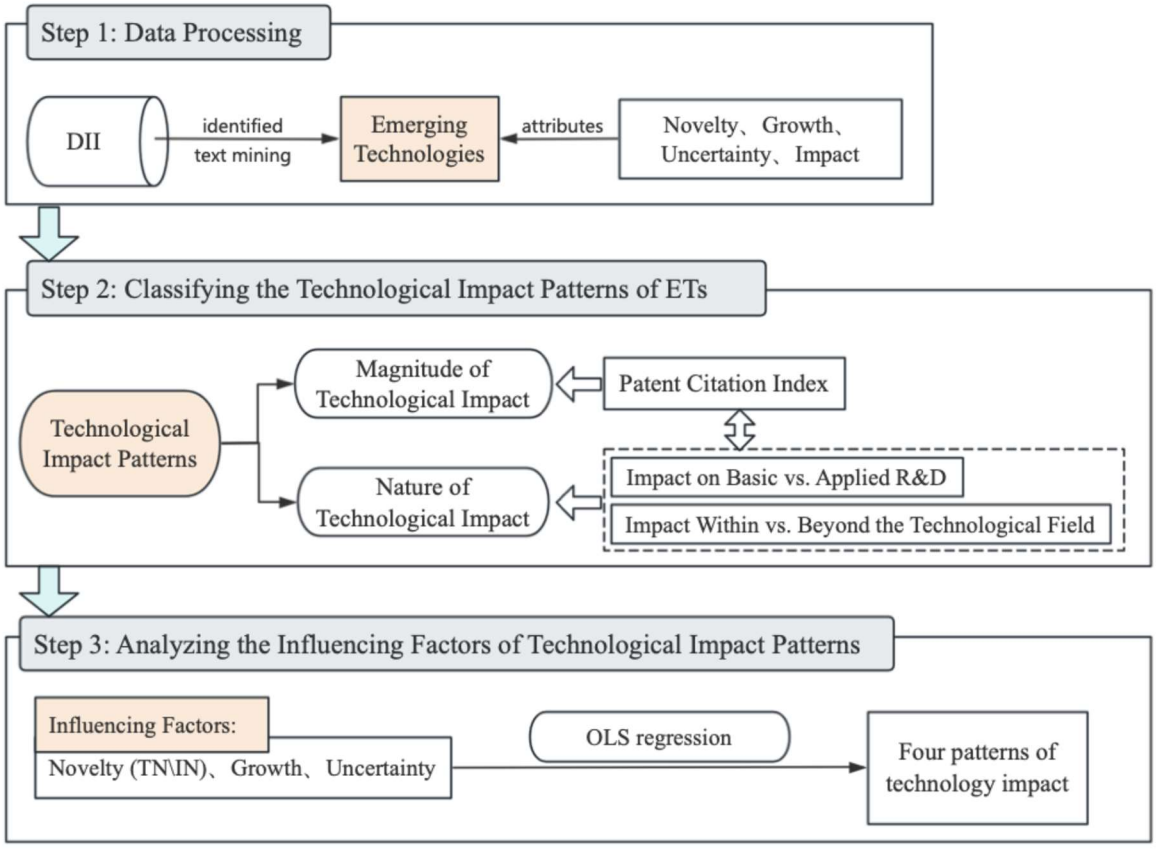


Figure 1. Research framework

Categorization of technological impact patterns

Magnitude of technological impact

Previous studies indicate that the frequency of direct patent citations is a viable metric for measuring technological impact (Glänzel & Thijs, 2012; Wang, 2018). The prominent impact of a term can be quantified by the standardized citation counts appearing in patent records. To account

for variances in term representations across patents, normalized TF-IDF values are utilized to weight terms accordingly. The prominent impact (PI) of a term is calculated as follows:

$$\mathrm{PI} = \sum \alpha_i C_{i,t} \qquad (1)$$

where $C_{i,t}$represents citation counts, and α_i denotes the normalized TF-IDF values for the term in the i-th patent record, aggregating the values across all records to derive the term's PI index.

Types of technological impact

Existing studies indicate that the manifestations of technological influence vary according to technological characteristics, application contexts, and knowledge dissemination pathways (Messeni Petruzzelli et al., 2015). To explore whether different modes of technological influence are affected by distinct factors, this paper categorizes technological influence into four patterns based on the scope and direction of impact: within the technological domain, outside the technological domain, basic research and development (R&D) influence, and applied R&D influence.

(1) Impact within vs. beyond the technological field

Emerging technologies in the digital medical sector often have extensive cross-disciplinary impacts. Analyzing how novelty, growth, and uncertainty influence different technology adoption paths aids in revealing the scope and trajectory of technology dissemination. It also helps determine whether a technology is more specialized or has interdisciplinary value. Therefore, we classify the impacts of emerging technologies based on their citations within or beyond their primary field. For instance, a technology is considered to have an internal field impact if it is cited by patents within the digital medical field and an external field impact if cited by patents in other fields. This categorization extends to calculating PI_In and PI_Out indices:

$$\mathrm{PI_In} = \sum \alpha_i Cin_{i,t} \qquad (2)$$

$$\mathrm{PI_Out} = \sum \alpha_i Cout_{i,t} \qquad (3)$$

where $Cin_{i,t}$ and $Cout_{i,t}$ denote the citation counts within and outside the digital medical technology field, respectively.

(2) Impact on basic vs. applied R&D

Existing research shows a propensity for companies to focus on applied research while universities lean towards basic research (Poege et al., 2019; Kwon et al., 2021). Therefore, we classify the impact based on the patent holder's type, such as a company or university. If the holder is a company, the impact is considered as applied R&D; if a university, then as basic R&D. The indices PI_Ba and PI_Ap are obtained as follows:

$$\mathrm{PI_Ba} = \sum \alpha_i Cu_{i,t} \qquad (4)$$

$$\mathrm{PI_Ap} = \sum \alpha_i Cc_{i,t} \qquad (5)$$

where $Cu_{i,t}$ and $Cc_{i,t}$ represent the citation counts by university and company patent holders within the digital medical technology field, respectively.

Factors influencing the impact patterns of emerging technologies

The attributes of emerging technologies co-evolve (Rotolo et al., 2015), and while quantitative evidence suggests that novelty is a significant factor influencing technological impact (Uzzi et al., 2013; Messeni Petruzzelli et al., 2015), qualitative evidence also points to growth and uncertainty as influential (Messeni Petruzzelli et al., 2011). However, a systematic quantitative study of how these

intrinsic attributes affect technological impact is still lacking. Following the classification of technological impact patterns outlined earlier, this study employs regression analysis to explore the effects of temporal novelty, value novelty, and uncertainty on different technological impact patterns.

Dependent variables

The dependent variables in this study are the measures of technological impact of emerging technologies, including the overall impact (PI), impact within the technological field (PI_In), impact beyond the technological field (PI_Out), impact on basic development (PI_Ba), and impact on applied development (PI_Ap).

Independent variables

The independent variables are other attributes of emerging technologies, including temporal novelty, value novelty, growth, and uncertainty:

Temporal novelty (TN) measures the recency of technology emergence, suggesting newer technologies with later emergence dates. It is quantified by the mean filing year of patents associated with a given term, as follows:

$$TN = \sum_{i=1}^{n} t_i \, / n$$

Here, t_i denotes the filing year of the patent term, and n is the number of patents containing the term.

Innovation novelty (IN) measures the degree of technological innovation from the perspective of technical content. Breitzman and Thomas (2015) highlight that the recombination of technical knowledge is a key source of novelty (Fleming & Sorenson, 2001; Strumsky & Lobo, 2015). A broader range of knowledge absorbed during recombination indicates higher novelty (Jiang et al., 2024; Savino et al., 2017). Thus, the IN of a term is defined as the diversity of technical fields reflected in its backward citations:

$$IN = N \sum filed_{Bx}$$

Here, $\sum filed_{Bx}$ denotes the set of technology fields from backward patents where the term i appears in the patent x. The variable N signifies the total number of patents incorporating term i, and the IN denotes the cumulative technical fields from all backward citations in these patents. This classification of technological fields is based on the WIPO Technology Concordance Table, which defines 35 standard fields (WIPO35) (Jiang et al., 2024).

Growth (R) is measured using the average growth rate. To mitigate random fluctuations in the annual number of patents, we follow the approach of Q. Wang (2018) by applying a smoothing technique, where the number of patents for a given year is calculated as the average over the preceding three years. The growth rate of a term is then defined as its average annual increase.:

$$\mathrm{R} = \mathrm{A}ve \sum_{i} \frac{p_t}{p_{t+1}}$$

Here, p_t denotes the number of ET term appears in patents records at time t.

Uncertainty (H) is measured using the H-index, originally proposed by Brillouin (1956) as a measure of uncertainty based on the principle of information entropy. This study applies the H-index to evaluate the uncertainty surrounding the future applications of emerging technologies (ETs) (Jiang et al., 2024). A higher H-index for a term reflects greater ambiguity in its functional definition and substantial inherent uncertainty in its potential applications.

$$H = \frac{lgN! - \sum(lgN_i!)}{N}$$

Here, N_i represents the occurrence frequency of a term within technical field i, and N represents the aggregate frequency of the term across all technical fields.

The theoretical basis and detailed explanation of these metrics are further elaborated in Jiang et al. (2024). Despite the linear relationships among dependent variables, such as the correlation between growth and novelty discussed in related work (Messeni Petruzzelli et al., 2015), these relationships do not compromise the investigation of technological impact as these attributes are inherent to emerging technologies and are likely interconnected (Rotolo et al., 2015).

Control variables

Several control variables were incorporated to account for variations in technical impact. Firstly, we controlled for the *number of patents* associated with emerging technology, as the volume of patents can influence the technological impact. Secondly, the size of the team involved in technology development was included as a control variable because economies of specialization, a larger and more diverse knowledge base, and a broader and more heterogeneous external network can influence the impact of the technology (Messeni Petruzzelli et al., 2015). Thus, the number of inventors involved in the emerging technology (*TeamSize*) was included (Messeni Petruzzelli et al., 2015). Thirdly, the *patent family size* was controlled for, as a larger patent family typically has a higher technological impact (Corredoira & Banerjee, 2015). Fourthly, we controlled for the distribution of patents across different *countries* (Kwon et al., 2021).

For the regression analysis, we fit our data to an OLS regression model using robust standard errors to take into account probable violation of the homoskedasticity assumption in estimation.

Analysis and results

Descriptive analysis

Table 1 displays the correlations among the key variables. All correlation coefficients are below 0.4, indicating that multicollinearity is not a significant concern in the regression analysis (Kwon et al., 2021). Furthermore, the examination of the variance inflation factor (VIF) associated with each regression coefficient, ranging from 1.02 to 1.16, confirms the absence of collinearity concerns, as values less than 10 suggest negligible collinearity (Messeni Petruzzelli et al., 2011). Below is a representation of the correlation matrix and VIF values extracted from the data:

	TN	IN	R	H	PI	VIF
TN	1					1.02
IN	-0.012	1				1.15
R	0.150	0.014	1			1.03
H	0.016	0.363	0.061	1		1.16
PI	-0.008	0.181	0.007	0.100	1	/

Table 1. Correlation between variables

Regression results

Table 2 displays the OLS estimation results for different types of technological impact. Model 1 provides an analysis of overall patent impact (Y_PI), while Models 2-5 test the technological impacts categorized by internal vs. external technological fields (PI_In vs. PI_Out) and by basic vs. applied R&D (PI_Ba vs. PI_Ap).

Model 1 highlights a U-shaped relationship between temporal novelty (TN) and technological impact. This nonlinear relationship is significant and robust, with a correlation coefficient as high as 0.6, indicating that although older technologies might generate greater impact due to the

cumulative effect of time, newer technologies are more likely to be identified by innovators to highlight technological progress. This is particularly true in emerging technology fields where newer technologies can achieve higher impacts. Innovation novelty (IN) consistently shows a positive relationship with technological impact, reinforcing the notion that more innovative technologies are likely to evolve into high-impact technologies. This finding aligns with prior research that posits innovation as a driver of technological impact (Corredoira & Banerjee, 2015; Kwon et al., 2021). However, the growth rate (R) and uncertainty (H) did not exhibit significant effects in most models, suggesting that the simple annual patent growth rate may not be a strong predictor of technological impact within this dataset.

The significant results for technological novelty are clear, while the non-significant effects of growth and uncertainty on technological impact are expected. Growth does not always positively influence impact. In early-stage technologies, small bases amplify base effects, leading to potentially negative correlations (Porter et al., 2019). High growth rates often reflect proportional accumulation rather than enhanced influence and may result from novelty-driven growth (Castaldi et al., 2015; Rosiello & Maleki, 2021). However, rapid growth can over-concentrate resources, intensify competition, and diminish impact in later stages due to resource constraints or market saturation (Corredoira & Banerjee, 2015).

Models 2 and 3 provide evidence that several impact factors differ significantly between technological impacts within the digital medical technology field and impacts outside of it. Specifically, it appears that TN, IN, R, and H show no significant effects within the technological field (PI_In), whereas for the technological impact outside of the field (PI_Out), TN exhibits a positive U-shaped relationship, and IN shows a positive linear relationship, as shown in Figure 2. This indicates that new technologies might initially struggle to make an impact within their primary fields but gain significant traction and influence over time in external fields. This may suggest that while a technology's direct field may be slow to recognize its utility or potential, auxiliary fields might quickly discover new applications or value.

Models 4 and 5 indicate that both TN2 and IN have a significant positive impact on applied research, with technological innovation novelty positively influencing both basic and applied research impacts, albeit more pronounced in applied research, as shown in Figure 3. This suggests that in the context of applied development, newer and more innovative technologies may have a greater impact. Here, the temporal novelty reflects a preference for actual applications, indicating a favour for newer innovative solutions.

These contrasts between internal vs. external impacts and basic vs. applied research impacts provide valuable insights into how different sectors and types of research within the digital medical technology field perceive and value technological innovations. The significant differences in how temporal and innovation novelties are valued across these models suggest that varying strategies may be necessary to maximize technology adoption and impact according to the specific target domain or research focus.

Variables	(1) Y_PI	(2) PI_In	(3) PI_Out	(4) PI_Ba	(5) PI_Ap
TN	-0.248***	-0.008	-0.240***	-0.236***	-0.018*
	(0.080)	(0.007)	(0.077)	(0.075)	(0.010)
TN2	0.615***	0.020	0.594***	0.585***	0.044*
	(0.199)	(0.018)	(0.191)	(0.187)	(0.026)
IN	0.072***	0.001	0.071***	0.067***	0.005**
	(0.019)	(0.002)	(0.019)	(0.018)	(0.003)
R	-0.004	-0.000	-0.004	-0.004	-0.001
	(0.009)	(0.001)	(0.009)	(0.009)	(0.001)
H	0.500	-0.131*	0.632	0.555	-0.130
	(0.789)	(0.071)	(0.758)	(0.742)	(0.103)
Patent_Num	0.000	0.000	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Team_Num	-0.016	0.001	-0.017*	-0.012	-0.002*
	(0.010)	(0.001)	(0.010)	(0.010)	(0.001)
PatentFamily_Num	0.113***	0.004**	0.109***	0.108***	0.005**
	(0.020)	(0.002)	(0.019)	(0.019)	(0.003)
Country_Num	-0.003***	-0.000***	-0.003***	-0.003***	-0.000**
	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)
_cons	2.5e+06***	8.3e+04	2.4e+06***	2.4e+06***	1.8e+05*
	(8.1e+05)	(7.3e+04)	(7.8e+05)	(7.6e+05)	(1.1e+05)
r2_a	0.065	0.006	0.069	0.069	0.009
N	1016	1016	1016	1016	1016

*p < 0.10; **p < 0.05; ***p < 0.01; Standard errors in parentheses

Table 2. OLS regression results for different types of technological impact

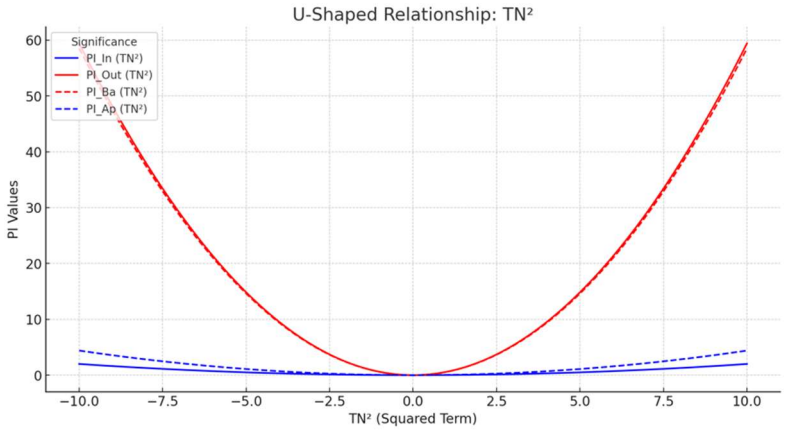


Figure 2. The U-shaped relationship between TN² and PI

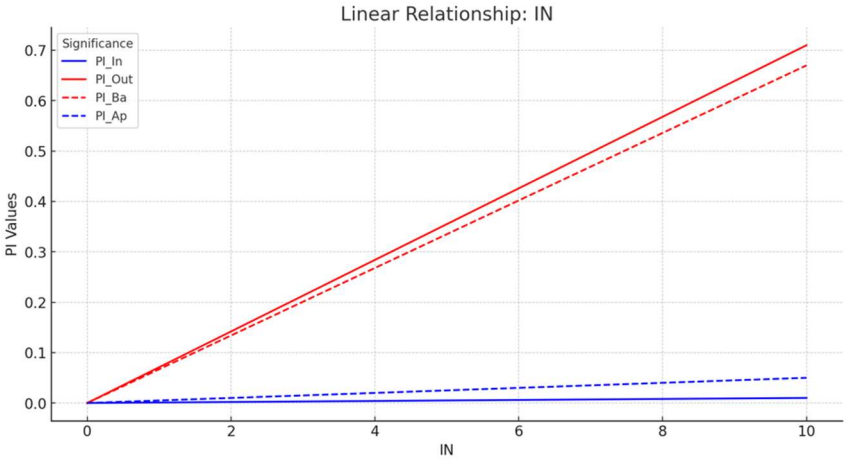


Figure 3. The linear relationship between IN and PI

Robustness testing

We performed robustness testing on the OLS results by applying a log transformation to the dependent variables. The results, consistent with Table 2. Additionally, fractional response regression models were used for further validation. Table 3 reports the outcomes, providing additional evidence for the relationship between technological novelty and technological impact.

Variables	lnY_PI	lnPI_In	lnPI_Out	lnPI_Ba	lnPI_Ap
TN	-1.4606***	-0.8811	-1.4903***	-1.3979*	-1.4910***
	(0.515)	(0.777)	(0.523)	(0.791)	(0.536)
TN2	1.6329***	0.9658	1.6636***	1.3845*	1.6739***
	(0.482)	(0.774)	(0.495)	(0.804)	(0.501)
IN	0.5540***	0.1343	0.5805***	0.4322**	0.5688***
	(0.147)	(0.210)	(0.146)	(0.217)	(0.154)
R5	-0.0793	0.1447	-0.0915	-0.4588	-0.0419
	(0.362)	(0.332)	(0.368)	(0.489)	(0.350)
H	0.1183	-0.3606*	0.1602	-0.1948	0.1520
	(0.147)	(0.196)	(0.149)	(0.187)	(0.156)
Patent_Num	0.1632	0.2582	0.1501	-0.0126	0.1871
	(0.198)	(0.246)	(0.209)	(0.444)	(0.189)
Team_Num	-0.2442*	0.2805	-0.2832**	-0.5846**	-0.2178
	(0.132)	(0.202)	(0.134)	(0.250)	(0.133)
PatentFamily_Num	0.9274***	0.5117**	0.9437***	0.5346**	0.9511***
	(0.184)	(0.237)	(0.185)	(0.263)	(0.188)
Country_Num	-0.3843***	-0.4985***	-0.3674***	-0.3793**	-0.3740***
	(0.129)	(0.184)	(0.131)	(0.175)	(0.134)
Constant	-1.7588***	-1.5630***	-1.8152***	-1.3767***	-1.8382***
	(0.135)	(0.183)	(0.139)	(0.175)	(0.142)
Obs.	1016	1016	1016	1016	1016

*p < 0.10; **p < 0.05; ***p < 0.01; Standard errors in parentheses

Table 3. Robustness testing results

Conclusions

This study has systematically explored the impact patterns and influencing factors of emerging technologies, with a specific focus on digital medical technology. The analysis employed an

extensive dataset comprising 15,116 patent records to scrutinize the influence of various attributes of emerging technologies—namely temporal and innovation novelty, growth, and uncertainty—on their technological impact across different patterns.

There are some key findings. ***Firstly***, temporal novelty and impact dynamics: The relationship between temporal novelty and technological impact exhibits a pronounced U-shaped curve. This pattern highlights the distinct characteristics of emerging technologies, which diverge from traditional technologies whose influence typically grows incrementally over time. Instead, emerging technologies demonstrate a strong impact early in their development cycle, underscoring the rapid initial uptake and the potential for early significant influence within their respective fields. ***Secondly***, innovation novelty as a driver of impact. Innovation novelty consistently exhibits a positive influence on technological impact, corroborating the hypothesis that more innovative technologies are more likely to evolve into high-impact innovations. This relationship holds true across different contexts of technological application, from internal to external fields and from basic to applied research. ***Thirdly***, the study differentiates the impacts within versus beyond the technological field and between basic and applied research. Findings reveal that newer technologies find it challenging to make an immediate impact within their primary domain but often gain substantial traction externally. Furthermore, applied research shows a greater receptiveness to innovative and novel technologies compared to basic research, which may have longer gestation periods for impact realization. ***Finally***, limited role of growth and uncertainty. Contrary to initial expectations, growth rates and uncertainty did not consistently influence technological impact, suggesting that these factors alone may not be reliable predictors of impact within the dataset explored. However, the nature of uncertainty and the challenges in accurately measuring this variable warrant further exploration. This area has seldom been comprehensively addressed in existing studies, pointing to a significant gap in the current understanding and measurement of technological uncertainty (Jiang et al., 2024; Xu et al., 2021).

In conclusion, this study enhances our understanding of how different attributes of emerging technologies influence their patterns of impact, providing critical insights that can help guide future technological developments, strategic industry positioning, and policy formulations in the rapidly evolving landscape of digital medical technology. Further research could expand these insights into other fields of emerging technology to validate the generalizability of the findings and refine our understanding of innovation dynamics across sectors.

Acknowledgements

This research was supported by the Max Planck Institute for Innovation and Competition during the author's visit. The author also acknowledges the funding support from the China Scholarship Council (CSC).

About the author(s)

Man Jiang is a Ph.D. candidate at the School of Information Management, Wuhan University. She conducted this research during her visit at the Max Planck Institute for Innovation and Competition and is currently a visiting researcher at the Technical University of Munich. Her research focuses on emerging technology identification and technology innovation management. She can be contacted at jiangman@whu.edu.cn

Siluo Yang is a Professor at the School of Information Management, Wuhan University. He received his Ph.D. from Wuhan University, and his research focuses on informetrics and scholarly evaluation. He can be contacted at 58605025@qq.com

Shuang Tao is a Lecturer at the School of Economics, Wuhan Polytechnic University. He received his Ph.D. from Huazhong University of Science and Technology, focusing on corporate economic management. He can be contacted at taos0118@163.com

References

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2018). An indicator of technical emergence. Scientometrics, 115(1), 35–49. https://doi.org/10.1007/s11192-018-2654-5

Cattani, G. (2005). Preadaptation, Firm Heterogeneity, and Technological Performance: A Study on the Evolution of Fiber Optics, 1970–1995. Organization Science, 16(6), 563–580. https://doi.org/10.1287/orsc.1050.0145

Corredoira, R. A., & Banerjee, P. M. (2015). Measuring patent's influence on technological evolution: A study of knowledge spanning and subsequent inventive activity. Research Policy, 44(2), 508–521. https://doi.org/10.1016/j.respol.2014.10.003

Corrocher, N., Malerba, F., & Montobbio, F. (2003). The emergence of new technologies in the ict field: Main actors, geographical distribution and knowledge sources. 33.

Day, G. S., & H, P. J. (2000). Wharton On Managing Emerging Technologies. Wiley.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. Science, 359(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Furukawa, T., Mori, K., Arino, K., Hayashi, K., & Shirakawa, N. (2015). Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions. Technological Forecasting and Social Change, 91, 280–294. https://doi.org/10.1016/j.techfore.2014.03.013

Gao, Q., Liang, Z., Wang, P., Hou, J., Chen, X., & Liu, M. (2021). Potential index: Revealing the future impact of research topics based on current knowledge networks. Journal of Informetrics, 15(3), 101165. https://doi.org/10.1016/j.joi.2021.101165

Glänzel, W., & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. Scientometrics, 91(2), 399–416. https://doi.org/10.1007/s11192-011-0591-7

Huang, H. (2020). The effect of the small-firm dominated ecology on regional innovation. The Annals of Regional Science, 65(3), 703–725. https://doi.org/10.1007/s00168-020-01000-7

Jiang, M., Yang, S., & Gao, Q. (2024). Multidimensional indicators to identify emerging technologies: Perspective of technological knowledge flow. Journal of Informetrics, 18(1), 101483. https://doi.org/10.1016/j.joi.2023.101483

Kwon, S., Youtie, J., & Porter, A. L. (2021). Interdisciplinary knowledge combinations and emerging technological topics: Implications for reducing uncertainties in research evaluation. Research Evaluation, 30(1), 127–140. https://doi.org/10.1093/reseval/rvaa029

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. Journal of the Association for Information Science and Technology, 66(7), 1323–1332. https://doi.org/10.1002/asi.23266

Lee, C., Cho, Y., Seol, H., & Park, Y. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. Technological Forecasting and Social Change, 79(1), 16–29. https://doi.org/10.1016/j.techfore.2011.06.009

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. Technological Forecasting and Social Change, 127, 291–303. https://doi.org/10.1016/j.techfore.2017.10.002

Lee, H., Lee, S., & Yoon, B. (2011). Technology clustering based on evolutionary patterns: The case of information and communications technologies. Technological Forecasting and Social Change, 78(6), 953–967. https://doi.org/10.1016/j.techfore.2011.02.002

Lee, W. H. (2008). How to identify emerging research fields using scientometrics: An example in the field of Information Security. Scientometrics, 76(3), 503–525. https://doi.org/10.1007/s11192-007-1898-2

Li, M., Porter, A. L., & Suominen, A. (2018). Insights into relationships between disruptive technology/innovation and emerging technology: A bibliometric perspective. Technological Forecasting and Social Change, 129, 285–296. https://doi.org/10.1016/j.techfore.2017.09.032

Litvinski, O. (2018). Emerging Technology: Toward a Conceptual Definition. International Journal of Trade, Economics and Finance, 9, 257–263. https://doi.org/10.18178/ijtef.2018.9.6.625

Messeni Petruzzelli, A., Maria Dangelico, R., Rotolo, D., & Albino, V. (2011). Organizational factors and technological features in the development of green innovations: Evidence from patent analysis. Innovation, 13(3), 291–310. https://doi.org/10.5172/impp.2011.13.3.291

Messeni Petruzzelli, A., Rotolo, D., & Albino, V. (2015). Determinants of patent citations in biotechnology: An analysis of patent influence across the industrial and organizational boundaries. Technological Forecasting and Social Change, 91, 208–221. https://doi.org/10.1016/j.techfore.2014.02.018

Nemet, G. F., & Johnson, E. (2012). Do important inventions benefit from knowledge originating in other technological domains? Research Policy, 41(1), 190–200. https://doi.org/10.1016/j.respol.2011.08.009

Noh, H., Song, Y.-K., & Lee, S. (2016). Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations. Telecommunications Policy, 40(10), 956–970. https://doi.org/10.1016/j.telpol.2016.04.003

Phene, A., Fladmoe-Lindquist, K., & Marsh, L. (2006). Breakthrough innovations in the U.S. biotechnology industry: The effects of technological space and geographic origin. Strategic Management Journal, 27(4), 369–388. https://doi.org/10.1002/smj.522

Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2019). Emergence scoring to identify frontier R&D topics and key players. Technological Forecasting and Social Change, 146, 628–643. https://doi.org/10.1016/j.techfore.2018.04.016

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? Research Policy, 44(10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. Research Policy, 43(8), 1450–1467. https://doi.org/10.1016/j.respol.2014.02.005

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. Science, 342(6157), 468–472. https://doi.org/10.1126/science.1240474

Wang, Q. (2018). A bibliometric model for identifying emerging research topics. Journal of the Association for Information Science and Technology, 69(2), 290–304. https://doi.org/10.1002/asi.23930

Xu, H., Winnink, J., Yue, Z., Zhang, H., & Pang, H. (2021). Multidimensional Scientometric indicators for the detection of emerging research topics. Technological Forecasting and Social Change, 163, 120490. https://doi.org/10.1016/j.techfore.2020.120490

Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. Journal of Informetrics, 13(4), 100983. https://doi.org/10.1016/j.joi.2019.100983

Xu, S., Hao, L., Yang, G., Lu, K., & An, X. (2021). A topic models based framework for detecting and forecasting emerging technologies. Technological Forecasting and Social Change, 162, 120366. https://doi.org/10.1016/j.techfore.2020.120366

Zhou, Y. (2019). A novel method to identify emerging technologies using a semi-supervised topic clustering model: A case of 3D printing industry. Scientometrics, 120, 167–185.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Leveraging scientific knowledge for technological inventions: insights from knowledge recombination theory

Guiyan Ou, Haodong Chen, Kaili Wang, and Jiang Wu

DOI: https://doi.org/10.47989/ir30iConf47218

Abstract

Introduction. Technological inventions are often built on top of previous technological developments and sometimes earlier scientific discoveries. It has been demonstrated that technologies integrating scientific knowledge exhibit superior performance in terms of their technological value compared to non-science-based inventions. However, the factors influencing the utilisation of scientific knowledge components during the technological invention process remain unclear.

Method. By analysing 1,685,970 utility patents granted by the USPTO from 2001 to 2010, this study identifies five characteristics-rooted in the perspectives of knowledge recombination and inventor team-that potentially affect the adoption of scientific knowledge in technological inventions.

Results. Both Logit and Tobit regression analyses reveal that recombination complexity and novelty significantly enhance the propensity and intensity of integrating scientific knowledge into technological inventions, whereas knowledge maturity exhibits an inverse relationship. The most critical determinant identified is the inventor team's experience with science-based inventions, which significantly positively impacts the reliance on scientific knowledge, whereas a broader technical knowledge base within the team tends to reduce the use of scientific components in inventions.

Conclusions. The findings of this study provide insights into the unique recombination processes underlying science-based technologies and offer guidance for optimising organisational and policy designs for innovation activities.

Introduction

Knowledge recombination has been recognized as one of the crucial elements in understanding technological innovation since Joseph Schumpeter's influential work in 1934. Based on knowledge recombination theory, technological innovation can typically be conceptualized as a process of searching and recombining existing technological knowledge elements to develop solutions to technical problems and generate new knowledge (Fleming and Sorenson, 2001; Fleming and Sorenson, 2004; Xiao et al., 2022). However, with the rising complexity of technological issues, the generation and evolution of technological innovation are no longer confined to a single domain of technological knowledge. Instead, they increasingly involve integration and development across multiple scientific and technological fields (Zhao et al., 2023). This trend is clearly reflected in patent citation patterns. Arora et al. (2023) observed a significant increase in the integration of scientific literature within patents, where the proportion citing scientific articles rose from 6% in 1980 to 30% in 2015, and the average number of citations per patent escalated from 0.1 to 4.4. Moreover, science-based innovations exhibit relative advantages in terms of technological impact. Studies by Poege et al. (2019) and Hohberger (2016) have demonstrated that science-based innovations surpass non-science-based ones in both technological value and the rate of technological diffusion.

While the significance of scientific knowledge as a catalyst for technological innovation is increasingly recognized, the transformation of scientific knowledge into tangible innovations remains a complex endeavor. This is attributed to the distinct philosophical logics that underlie science and technology. The essence of science is the accurate discovery of natural phenomena or laws, usually representing more abstract principles touching upon the fundamental relations between causes and effects, but this may not always be the key to solving practical problems (Anckaert et al., 2020; Arora and Gambardella, 1994). Instead, technical knowledge embodies technicality and more contextualized content that stems from simulating real-world situations in experimental settings, making it easier to be incorporated into practical applications (Bozeman, 2000; Marx and Fuegi, 2020; Novelli, 2015). Given the significant disparities between scientific research activities and technological innovation activities, the academic community still lacks a deep understanding and systematic research on how scientific knowledge functions as an effective component in the process of technological invention. Therefore, it is essential to elucidate the mechanisms by which scientific knowledge is utilized in new technological developments. This not only deepens our understanding of its role in driving technological innovation but also equips practitioners with the insights necessary for more effective utilization of this knowledge in technological development.

As highlighted earlier, knowledge recombination is a mature structure that aids in comprehending the origins of technological innovation. Our study aims to explore the mechanisms underlying the use of scientific knowledge in technological inventions by focusing on the process of knowledge recombination during technological invention. The outcomes of technological innovation are contingent upon both the nature of the search scope and the capabilities of the inventors (Hur and Oh, 2021). Accordingly, this study delves into two primary dimensions: the characteristics of knowledge recombination and the traits of the inventor teams involved in technological invention. The research specifically addresses two critical questions:

(1) How do the features of knowledge recombination and the characteristics of the invention team influence the tendency to integrate scientific knowledge into technological inventions?

(2) How do these same factors affect the degree of scientific knowledge integration into technological inventions?

Related work and hypotheses

Knowledge recombination and scientific knowledge

Related studies have demonstrated that the characteristics of both the search space and the knowledge elements significantly impact the process of invention, resulting in variations in the outcomes of inventions (Savino et al., 2017; Fleming and Sorenson, 2001). This study first assesses the impact of three key characteristics associated with the search space—recombination complexity, novelty, and knowledge maturity—on the integration of scientific knowledge into technological invention development.

Recombination complexity

Prior research has shown that the size of the knowledge repository and its interdependencies (i.e., complexity) significantly affect the likelihood of successful knowledge recombination (Kauffman, 1993; Fusillo, 2023; Fleming and Sorenson, 2001). Excessive redundancy in the components of technological knowledge may restrict the range of potential knowledge exploration, inevitably leading to cognitive lock-in (Nooteboom, 2000). In contrast, the diversity of a knowledge base is regarded as a fundamental requirement for long-term technological development (Saviotti, 1996). A series of studies have shown that the expansion of technological knowledge components across different domains within a knowledge base not only broadens the scope of searchable knowledge but also generates more potential technological combinations, thereby increasing the likelihood of successful recombination (Rosenkopf and Nerkar, 2001; Novelli, 2015; Baumann and Siggelkow, 2013). However, greater diversity in the knowledge base also increases the complexity of technological interdependencies. The invention process not only requires managing the substantial noise information resulting from the expanded scope of knowledge search while avoiding the trap of local optima (Dunne and Dougherty, 2016). Choi et al. (2018) observe that scientific knowledge facilitates the comprehension of interdependencies among technologies, the identification of effective technological combinations, and the enhancement of efficiency in complex knowledge recombination. It can be inferred that the more complex the process of knowledge recombination, the greater the likelihood that scientific knowledge being utilized in the invention process. Based on this, we propose the following hypotheses:

H1a: The higher the complexity of knowledge recombination, the greater the likelihood of incorporating scientific knowledge components in the invention process.

H1b: The higher the complexity of knowledge recombination, the more scientific knowledge components are integrated during the invention process.

Recombination novelty

Recombining technological knowledge components from disparate domains results in different levels of novelty. Keijl et al. (2016) highlights that inventions derived exclusively from the recombination of knowledge components within local technological fields typically demonstrate lower levels of novelty. Conversely, technological combinations based on distant technological origins tend to exhibit higher degrees of novelty. Related empirical evidence supports the idea that knowledge recombinations exhibiting higher novelty possess greater potential value for future technological development. For example, Kaplan and Vakili (2015), based on an analysis of 2,826 granted patents in the nanotube sector, found that novel knowledge recombinations are not only more likely to form the basis for subsequent technological advancements but also yield greater economic value. However, high novelty knowledge recombinations often come at the cost of reduced familiarity and recognizability (Keijl et al., 2016; Aldrich and Fiol, 1994). Due to limited familiarity and recognizability of technological knowledge components, elucidating the vast array of potential technological combinations and their potential impacts on technical issues proves challenging during extensive searches (Afuah and Tucci, 2012). Moreover, distant technological knowledge components may not provide readily usable building blocks. Therefore, to mitigate the

variability and uncertainty inherent in knowledge recombination due to the pursuit of novelty, the invention process is likely to depend more heavily on scientific knowledge. On one hand, scientific reasoning helps inventors identify invention opportunities that are far removed from local technological fields. On the other hand, the predictive function of scientific knowledge can better ascertain potential technological alternatives, thereby avoiding experiments doomed to fail (Balconi et al., 2010). Based on this, this study formulates the following hypotheses:

H2a: The higher the novelty of knowledge recombination, the greater the likelihood of incorporating scientific knowledge components in the invention process.

H2b: The higher the novelty of knowledge recombination, the more scientific knowledge components are integrated during the invention process.

Knowledge maturity

The foundation of inventive activities may rest upon knowledge bases developed at different times, encompassing both recent and longstanding knowledge. The maturity level of the technological knowledge elements used during the recombination of knowledge can influence the degree of dependence of the invention on established scientific knowledge. Utilizing relatively mature (i.e., older) knowledge enables inventors to gain enhanced insights into the technical attributes and potential uses from industry-specific sources. This facilitates effective evaluation of the merits and demerits of knowledge during the recombination process, thereby reducing the likelihood of technological misapplication and errors (Turner et al., 2013). However, incorporating mature knowledge elements, could result in outdated technological solutions, given that numerous potential combinations of technology might have been previously utilized, thereby leaving less room for valuable technological innovations (Sørensen and Stuart, 2000). In contrast, recent knowledge is well-suited to meet changing requirements, reducing the risk of falling into capability traps and increasing the opportunities for knowledge recombination, which supports the ongoing effectiveness of innovation (Petruzzelli et al., 2018; Levinthal and March, 1993; Ahuja and Morris Lampert, 2001). Nevertheless, compared to well-established knowledge that has been validated over time, the benefits of integrating recent knowledge may be offset by technological uncertainties and higher implementation costs (Heeley and Jacobson, 2008; Petruzzelli and Savino, 2014). Nightingale (1998) notes that scientific knowledge is not merely a tool for generating answers but a means to comprehend how technology functions. This understanding can reduce technological uncertainties and lower the costs of trial and error. Consequently, to maximize the benefits of integrating recent technological knowledge, knowledge recombination may increasingly rely on scientific knowledge. This study proposes the following hypotheses:

H3a: The lower the maturity of the recombined technological knowledge, the greater the likelihood of incorporating scientific knowledge components in the invention process.

H3b: The lower the maturity of the recombined technological knowledge, the more scientific knowledge components are integrated during the invention process.

Inventor team and scientific knowledge

As active contributors to the process of invention, the attributes of inventors, including their knowledge and experience, play a critical role in shaping both the process of knowledge search and the outcomes of knowledge recombination within these inventive activities (Perry-Smith and Shalley, 2014; Ardito et al., 2016). Nevertheless, it is not yet known to what extent and in what ways the attributes of inventor teams impact their inclination or level of utilization of scientific knowledge components during the process of invention. The study primarily investigates two significant characteristics of inventor team: experience in science-based invention and breadth of technical knowledge.

Experience in science-based invention

Inventive experience encompasses the skills, knowledge, and resources that inventors accumulate through prior activities. Existing research suggests that the inertia of past experiences influences the approaches inventors take to seek technical solutions (Liao et al., 2008). In other words, for invention teams with extensive inventive experience, the inertia of past experiences drives them to create new technological combinations based on their existing knowledge base, rather than seeking new knowledge outside their current field (Wang et al., 2017). Consequently, from this perspective, an inventor's past experience with science-based inventions enhances the likelihood of utilizing scientific knowledge to address new technological challenges in the future. Furthermore, while inventors may draw upon the accumulated knowledge produced by technological progress, a thorough and accurate understanding of science enables them to search and recombine information effectively (Ardito et al., 2021). Previous successful experiences in the application of scientific knowledge by inventors have been shown to ease the challenges associated with integrating new scientific and technological insights, thereby reducing the time spent on knowledge exploration (Novelli, 2015). Based on this, the following hypotheses are proposed in this study:

H4a: The richer the science-based invention experience of the inventor team, the higher the likelihood of incorporating scientific knowledge components during the invention process.

H4b: The richer the science-based invention experience of the inventor team, the greater the number of scientific knowledge components they incorporate during the invention process.

Breadth of technological knowledge

The breadth of an inventor's technological knowledge can be defined as all technological knowledge elements they possess, reflecting the diversity of the technological knowledge accumulated (Fleming et al., 2007). Simonton (2003) demonstrated that a rich diversity of cognitive elements available for association enhances the likelihood of identifying unique combinations within these elements. This implies that the broader the scope of technological knowledge, the greater the number of options available to an inventor when addressing new problems. In an investigation targeting inventors in the 3M innovation domain, Boh et al. (2014) employed a mixed-method approach combining qualitative (case study) and quantitative research to investigate how the breadth of inventors' expertise influences their innovative outcomes. Results indicate that extensive professional knowledge indeed plays a crucial role in enhancing inventors' capacity to produce more inventions. Furthermore, during the process of knowledge recombination, inventors may opt to utilize their existing knowledge base or seek out new external information to formulate solutions for distinct technological problems. Wang et al. (2017) pointed out that inventors typically initiate their process by exploring the existing knowledge space, employing iterative trial-and-error methods to test various combinations of existing knowledge until a satisfactory solution is found. Therefore, we argue that the broader the technological knowledge possessed by inventors or invention teams, the less likely they are to seek external scientific knowledge. This study proposes the following hypothesis:

H5a: The broader the range of technical knowledge possessed by an inventor team, the lower the likelihood of incorporating scientific knowledge components during the invention process.

H5b: The broader the range of technical knowledge possessed by an inventor team, the fewer the number of scientific knowledge components they incorporate during the invention process.

Measurement indicators

Knowledge recombination characteristics

Recombination complexity (*Complexity*). Complexity reflects the diversity of the knowledge base and components required in the process of knowledge recombination (Barbieri et al., 2020). This

study employs the Originality index proposed by Trajtenberg et al. (1997) as a proxy for recombination complexity. The formula for its calculation is as follows:

$$Originality_p = 1 - \sum_k^{N_p} \left(\frac{Nreferences_{pk}}{Nreferences_p} \right)^2 \quad (1)$$

Where$k$represents a specific technological field, and$N_p$denotes the total number of distinct technological fields represented in all patent references cited by patent$p$. This study uses the first four digits of the IPC classification (hereafter referred to as 'IPC4') to represent technological fields.$Nreferences_p$ indicates the number of patent references cited by patent$p$, and $Nreferences_{pk}$ represents the number of those references that belong to the technological field k. The Originality index reflects the degree to which the cited references of a patent are distributed across various technological fields, with values ranging from 0 to 1. A higher originality index indicates that the focal patent draws from a wide variety of technological domains, suggesting that it is the result of combining knowledge from multiple fields.

Recombination novelty (*Novelty*). Novelty represents the uniqueness in the process of knowledge recombination, typically reflecting the '*distance*' between new technologies and prior technological knowledge (Barbieri et al., 2020). This study utilizes the improved radicalness index by Squicciarini et al. (2013) to measure novelty. This index not only characterizes the knowledge gap between the technological categories of the focal patent and those of the patents it cites, but also reflects the extent to which the technological invention combines knowledge components in a novel manner. The formula is as follows:

$$Radicalness_p = \sum_j^{n_p} \frac{CT_j}{n_p}; IPC_{pj} \neq IPC_p \ (2)$$

Where $n_p$denotes the total number of IPC classification codes in all patent documents cited by the focal patent p. CT_j is the number of IPC4 codes in the cited patent$j$ that differ from those of the focal patent p. This index ranges from 0 to 1, where a higher value signifies greater novelty of the patent.

Knowledge maturity (*Maturity*). Following the approaches of Jiao (2022), we use the average temporal distance between focal patents and backward citations as an indicator of knowledge maturity. The specific formula used is presented below.

$$Maturity = \frac{\sum_{i=1}^{N} (Year_{focal} - Year_i)}{N} \quad (3)$$

Where N is defined as the number of backward citations to the focal patent, $Year_{focal}$ represents the application year of the focal patent, and $Year_i$ refers to the application year of the ith patent cited by the focal patent.

Inventor team characteristics

Experience in science-based invention (*Inventor_Sci_Expe*). This study quantifies the experience in science-based invention of the inventor team by calculating the number of patents citing scientific papers among those previously granted to the inventors. Specifically, for the focal patent p, this study tracks all patents granted to each inventor prior to the applications date of patent p, counts the number of patents that cite scientific literature for each inventor, and uses the maximum value to represent the team's experience in science-based invention for patent p.

Breadth of technological knowledge (*Inventor_Tech_Bread*). To investigate the breadth of technological knowledge utilized by the inventor team, this study calculates the total number of distinct IPC4 codes among the patents granted to the inventor team of focal patent p over the past five years, following the method proposed by Chen et al. (2021).

$$Technical\ Knowledge_p = n_p; n \in \{IPC4_1; \ldots IPC4_i; IPC4_j; \ldots; IPC4_n\}\ \&\ IPC4_i \neq IPC4_j \quad (4)$$

Data, variables, and models

Data collection

To test the hypotheses proposed in this study, we construct a patent-level dataset including U.S. utility patents granted between 2001 and 2010. The dataset consists of two primary datasets:(i) the granted patents research dataset from PatentsView and (ii) the '*reliance on science*' dataset, which provides pairs of citing patents and cited scientific publications by matching U.S. patents to publications in Microsoft Academic Graph with various degrees of confidence (March, 1991). In our study, we match our patent sample to scientific papers cited in patents' front page or in-text with a confidence score of at least 7. After filtering out records with incomplete information, our research sample includes 1,685,970 patents, of which 463,393 cite scientific articles. Overall, the dataset consists of these patents and 1,407,439 scientific publications, yielding a total of 6,473,214 unique citation pairs between patents and papers.

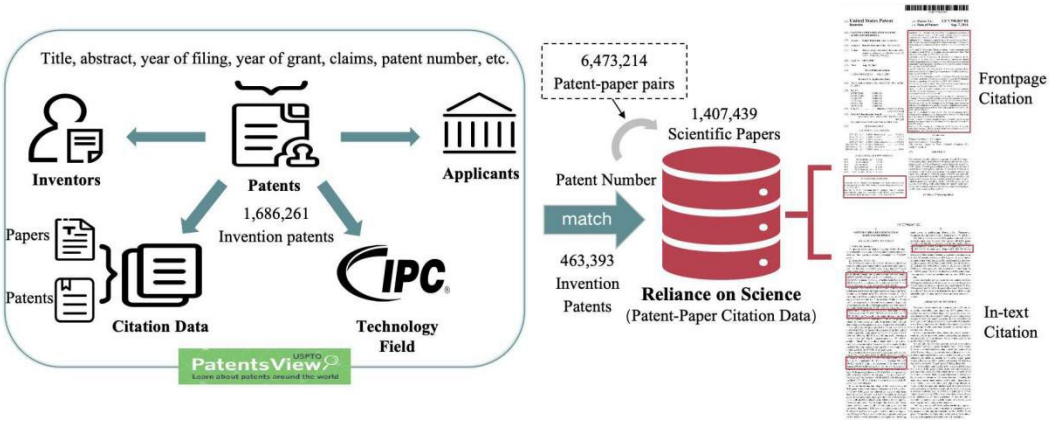


Figure 1. Flowchart of Data process

Variables

Dependent variables

This research delineates two dependent variables in accordance with the posed research questions: ***scientific knowledge integration tendency*** and ***scientific knowledge integration intensity***.

Scientific knowledge integration tendency (*Science*) assesses the tendency to integrate scientific knowledge into the technological development process, as measured by the citation of scientific papers by patents. In our study, a patent that cites at least one scientific paper, defined as a '*science-based invention*', is coded as 1. Conversely, a patent that does not cite any scientific papers, defined as a 'non-science-based invention,' is coded as 0.

Scientific knowledge integration intensity (*science_intensity*) refers to the degree to which scientific knowledge contributes to technological patent, as measured by the number of scientific papers cited by patents, using methodologies established in prior research (Harhoff et al., 2003; Sung et al., 2015).

Control variables

Taking into account potential influencing factors, this study incorporates a set of control variables, including the number of inventors (*Num_Inventors*), the number of applicants (*Num_Applicants*), applicant experience with science-based inventions (*Num_Applicants_Sci*), the number of backward patent citations (*Num_PC*), the speed of technology accumulation (*Tech_Cum_Speed*), and the type of applicant.

The variable *Num_Applicants_Sci* is incorporated because existing studies indicate that the knowledge base of patent applicants influences the trajectory and efficacy of their future technological innovations (Katila adn Ahuja, 2002; Maddala GS, 1987). We measured this by calculating the average number of patents granted that had cited scientific literature prior to the application.

Num_PC represents the count of patent documents that an invention's patent cites. By controlling for the number of backward citations associated with each invention, we aim to mitigate the extent of technological knowledge dependency, as discussed by Gruber et al. (2013).

Tech_Cum_Speed represents the speed at which a certain technology develops or evolves along its technological trajectory (Lee and Lim, 2001; Guo et al., 2013). The extent of cumulative development across technologies may affect the propensity of that the knowledge reorganization process relies on scientific knowledge (Grupp and Schmoch, 1992; Meyer, 2000). The formula is as follows:

$$Tech_Cum_Speed_p = \sqrt[5]{\frac{n_p}{n_5}} * 100\% \qquad (5)$$

Where n_p represents the cumulative number of the IPC4 classification codes for the focal patent p in its filing year. n_5 indicates the cumulative number of IPC classification codes from five years prior. The greater $Tech_Cum_Speed_p$ is, the faster the rate of accumulation in the technical field to which the invention belongs.

Regarding the type of applicant, referring to the applicant categorization algorithm proposed by Du Plessis et al. (2010), we classify patents into five categories: individual, company, government, university, and hospital, and set five dummy variables.

In addition, we include a set of time, country, and technology field dummies to control for potential time-varying effects, geographical and technological heterogeneity. Specifically, *Time* is a set of dummy variables for the patent's granted year. *Geo* denotes geographical dummies represented by the country of the first inventor listed on a patent. We also include the technological field dummies, *IPC*, at the 1-digits IPC level (The IPC is a hierarchical system and is divided into sections, classes, subclasses, main groups, and subgroups. The section of IPC includes A, B, C, D, E, F, G, and H, corresponding to eight broad technical fields).

Table 1 presents the descriptive statistics for each variable. The average Science is 0.275, indicating that 27.5% of patents in this study's sample cited at least one scientific literature. The average scientific_intensity is 3.839, with a maximum value of 3048, representing the highest number of citations to scientific literature by any single patent.

Variables	Obs	Mean	S.D.	Min	Max
Science	1685970	0.275	0.446	0	1
Scientific_Intensity	1685970	3.839	17.620	0	3048
Complexity	1685970	0.477	0.284	0	0.975
Novelty	1632772	0.474	0.341	0	1
Maturity	1660462	3359.372	3232.240	0	638550
Inventor_Sci_Expe	1685970	0.312	0.324	0	1
Inventor_Tech_Bread	1685970	5.734	8.076	0	148
Num_Inventors	1685970	2.510	1.804	0	76
Num_Applicants	1530794	1.034	0.213	1	14
Num_Applicants_Sci	1530794	2188.774	4688.280	0	45899
Num_PC	1685970	22.541	46.978	0	2399
Tech_Cum_Speed	1682142	1.047	0.093	0.574	2.887
Company	1530794	0.942	0.235	0	1
Government	1530794	0.025	0.157	0	1
University	1530794	0.027	0.163	0	1
Hospital	1530794	0.002	0.039	0	1
Individual	1530794	0.004	0.065	0	1

Table 1. Descriptive statistics for the main variables

Regression model

Given that the dependent variable, *science*, is binary, this study employs a Logit regression model for the primary regression analysis. Subsequent analyses utilize the Probit model to further verify the robustness of the results. The baseline Logit model is established as follows:

$$Logit(p) = ln(\frac{p}{1-p}) = \alpha + \beta_i x_i + \gamma_i Controls_i + Time_i + Geo_i + IPC_i + \varepsilon_i \quad (6)$$

Where x_i are the explanatory variables of this study, i.e., *Complexity*, *Novelty*, *Maturity*, *Inventor_Sci_Expe*, and *Inventor_Tech_Bread*. $Controls_i$ are the control variables of this study, $Time_i$, Geo_i, IPC_i denote the time, geography, and technology domain dummy variables, respectively.

The second dependent variable, *science intensity*, exhibits a significant number of zero entries, primarily because the majority of the samples represent '*non-science-based inventions*' that lack any scientific literature citations. To address potential biases that conventional count models might introduce, this research employs a Tobit regression model for analysis. The model is as follows:

$$Science_Intensity = \alpha + \beta_i x_i + \gamma_i Controls_i + Time_i + Geo_i + IPC_i + \varepsilon_i \quad (7)$$

Results

Logit regression results

This section investigates the impact of five characteristics on the variable 'Science' through logit regression analysis. Prior to the analysis, certain variables were logarithmically transformed to mitigate kurtosis and skewness, and multicollinearity was evaluated. Table 2 presents Spearman's correlation coefficients and the results of the multicollinearity tests, demonstrating low correlations (all below 0.7) and a maximum VIF of 2.05, confirming the absence of multicollinearity.

variables	VIF	1	2	3	4	5	6	7	8	9	10	11	12	13
1.Complexity	2.05	1												
2.Novelty	1.81	0.655	1											
3.Ln (Maturity)	1.20	0.138	0.097	1										
4.Inventor_Sci_Expe	1.19	0.059	0.010	-0.155	1									
5.Ln (Inventor_Tech_Bread)	1.43	0.070	0.050	-0.141	0.481	1								
6.Ln (Num_Inventors)	1.21	0.032	0.007	-0.108	0.218	0.432	1							
7.Ln (Num_Applicants)	1.03	0.007	0.007	-0.009	0.009	0.058	0.102	1						
8.Ln (Num_Applic_Sci)	1.21	-0.097	-0.079	-0.281	0.215	0.291	0.122	0.039	1					
9.Ln (Num_PC)	1.30	0.288	0.091	0.220	0.131	0.066	0.067	-0.026	-0.066	1				
10.Tech_Cum_Speed	1.06	-0.048	-0.010	-0.322	0.048	0.019	0.009	-0.021	0.171	-0.047	1			
11.Government	1.02	0.023	0.022	0.026	-0.012	-0.056	0.012	0.062	-0.050	-0.046	-0.029	1		
12.University	1.03	0.058	0.043	-0.012	0.065	-0.033	0.038	0.056	-0.002	-0.037	-0.043	-0.027	1	
13.Hospital	1.00	0.017	0.012	-0.009	0.027	-0.010	0.001	0.022	-0.008	-0.009	-0.016	-0.006	-0.007	1

Table 2. Spearman's correlation analysis and multicollinearity Test

The results of Logit regression analysis are shown in Table 3. Model (1) incorporates only control variables, while models (2) and (3) build upon model (1) by adding characteristics of knowledge recombination (i.e., *Complexity*, *Novelty and Maturity*) and characteristics of the inventor team (i.e., *Inventor_Sci_Expe*, *Invent_Tech_Bread*), respectively. Model (4) includes all explanatory and control variables. The full model demonstrates good fit with an AUC of 0.8159 and a classification accuracy of 78.52%. The last column of Table 3 details the marginal effects of the five explanatory variables included in Model (4).

Analysis spanning models 2 through 4 demonstrates consistent significance and directionality in the explanatory variables of interest. The findings from model 4 indicate that *Complexity* (β=0.221, $p<0.001$) and *Novelty* (β=0.120, $p<0.001$) are positively associated with *science*, confirming hypotheses H1a and H2a. Specifically, for each incremental increase in these recombination characteristics, the probability of incorporating scientific knowledge into the technology development process increases by 3.3% and 1.8%, respectively. In contrast, a significant negative relationship exists between *Maturity* (β=−0.345, $p<0.001$) and *science*, suggesting that less mature reorganized knowledge in technological inventions is more likely to incorporate scientific components. Moreover, a 1% increase in *Maturity* decreases the likelihood of using scientific knowledge in the invention process by 5.1%, thereby lending support to hypothesis H3a.

Furthermore, the results from Model 4 indicate a significant positive association between *Inventor_Sci_Expe* (β=2.211, $p<0.001$) and *science*, affirming hypothesis H4a. Each incremental unit of *Inventor_Sci_Expe* enhances the probability of utilizing scientific knowledge components by 32.6%. This finding underscores the role of experiential inertia, where inventors with extensive prior use of scientific knowledge are more inclined to integrate such elements into subsequent technological inventions. In contrast, *Invent_Tech_Bread* (β=−0.183, $p<0.001$) exhibits a negative correlation with *science*, where a 1% increase in *Invent_Tech_Bread* leads to a 2.7% decrease in the probability of employing scientific knowledge, supporting hypothesis H5a.

Variables	Model (1)	Model (2)	Model (3)	Model (4)	Average marginal effects
Complexity		0.192***		0.221***	0.033***
		(0.011)		(0.011)	(0.002)
Novelty		0.109***		0.120***	0.018***
		(0.008)		(0.009)	(0.001)
Ln (Maturity)		-0.374***		-0.345***	-0.051***
		(0.003)		(0.003)	(0.000)
Inventor_Sci_Expe			2.246***	2.211***	0.326***
			(0.007)	(0.007)	(0.001)
Ln (Invent_Tech_Bread)			-0.178***	-0.183***	-0.027***
			(0.003)	(0.003)	(0.000)
Control variables					
Ln (Nun_Inventors)	0.348***	0.315***	0.449***	0.421***	-
	(0.003)	(0.003)	(0.004)	(0.004)	
Ln (Num_Applicants)	0.250***	0.239***	0.295***	0.281***	-
	(0.016)	(0.016)	(0.017)	(0.017)	
Ln (Applicants_Sci)	0.064***	0.055***	0.035***	0.029***	-
	(0.001)	(0.001)	(0.001)	(0.001)	
Ln (Num_PC)	0.302***	0.410***	0.238***	0.330***	-
	(0.002)	(0.002)	(0.002)	(0.003)	
Tech_Cum_Speed	0.837***	0.487***	0.865***	0.545***	-
	(0.021)	(0.021)	(0.022)	(0.022)	
Government	0.947***	0.972***	0.982***	0.995***	-
	(0.012)	(0.012)	(0.012)	(0.013)	
University	2.149***	2.166***	2.109***	2.109***	-
	(0.014)	(0.014)	(0.014)	(0.015)	
Hospital	2.786***	2.712***	2.641***	2.534***	-
	(0.081)	(0.083)	(0.083)	(0.086)	
Individual	-0.254***	-0.172***	-0.162***	-0.097***	-
	(0.035)	(0.036)	(0.036)	(0.038)	

Note：Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001

Table 3. Logit regression results for *science*

Tobit regression results

In this section, the impact of five key explanatory variables on *science_intensity* was examined using Tobit regression, as detailed in Table 4. Model (5) served as the baseline with only control variables. Models (6) and (7) sequentially introduced additional variables pertaining to knowledge recombination and inventor team, based on the foundation established by model (5). Model (8) included all explanatory and control variables. The full model demonstrated an acceptable fit (pseudo R^2 = 0.176, Log likelihood = -1,238,628.6), although the normality assumption of error terms was violated (CM test: $p < 0.001$). We address this limitation through negative binomial regression in our robustness tests. The significance and direction of the five key explanatory variables remained consistent from model (6) to (8), indicating the robustness of the findings.

Firstly, the results from Model (8) showed that *Complexity* (β=0.306, p<0.001) and *Novelty* (β=0.148, p<0.001) were significantly and positively correlated with *science_intensity*. These results supported hypotheses H1b and H2b, suggesting that increased complexity and novelty in an invention's recombination are linked to a greater reliance on scientific knowledge. In contrast, *Maturity* (β=−0.413, p<0.001) has a significant negative effect on *science_intensity*. This indicates that an invention's knowledge maturity is inversely related to its dependence on scientific knowledge. Specifically, inventions developed within newer technological fields tend to incorporate more scientific components. Thus, hypothesis H3b is confirmed.

Secondly, *Inventor_Sci_Expe* (β=2.471, p<0.001) significantly positively influences *science_intensity*, whereas Invent_Tech_Bread (β=−0.211, p<0.001) demonstrates a significant negative effect. This suggests that a team with more experience in scientific invention tends to rely more heavily on scientific knowledge during technological development. Conversely, a broader range of technical knowledge among inventors correlates with less reliance on scientific knowledge components. Therefore, both H4b and H5b are supported.

Variables	**model（5）**	**model（6）**	**model（7）**	**model（8）**
Complexity		0.296*** (0.012)		0.306*** (0.011)
Novelty		0.148*** (0.009)		0.148*** (0.008)
Ln (Maturity)		-0.477*** (0.003)		-0.413*** (0.003)
Inventor_Sci_Expe			2.538*** (0.007)	2.471*** (0.007)
Ln (Invent_Tech_Bread)			-0.206*** (0.003)	-0.211*** (0.003)
Control variables				
Ln (Nun_Inventors)	0.433*** (0.004)	0.383*** (0.004)	0.499*** (0.004)	0.461*** (0.004)
Ln (Num_Applicants)	0.446*** (0.017)	0.427*** (0.017)	0.460*** (0.016)	0.439*** (0.016)
Ln (Applicants_Sci)	0.067*** (0.001)	0.055*** (0.001)	0.027*** (0.001)	0.019*** (0.001)
Ln (Num_PC)	0.374*** (0.002)	0.498*** (0.003)	0.268*** (0.001)	0.369*** (0.002)
Tech_Cum_Speed	0.835*** (0.023)	0.413*** (0.023)	0.800*** (0.023)	0.435*** (0.022)
Government	1.182*** (0.013)	1.181*** (0.013)	1.093*** (0.012)	1.082*** (0.012)
University	2.392*** (0.011)	2.371*** (0.012)	2.122*** (0.011)	2.094*** (0.011)
Hospital	3.036*** (0.046)	2.924*** (0.047)	2.564*** (0.042)	2.451*** (0.044)
Individual	-0.349*** (0.038)	-0.260*** (0.039)	-0.239*** (0.036)	-0.168*** (0.037)
Year	control	control	control	control
Country	control	control	control	control
IPC	control	control	control	control
Constant term	-3.046*** (0.051)	0.762*** (0.058)	-3.174*** (0.047)	0.130*** (0.054)
Pseudo R2	0.127	0.135	0.170	0.176
Observed Values	1,518,530	1,464,685	1,518,530	1,464,685

Table 4. Tobit regression results for *science_intensity*

Robustness test

To ensure the reliability and robustness of the research findings, this study conducted a revalidation of the empirical results by substituting both the model and the dependent variables. The outcomes confirmed that the direction and significance of all variables were consistent with the primary analysis results, thereby affirming the robustness of the findings. Detailed displays of these results are omitted here due to space limitations, but the robustness results are available upon request from the author.

Conclusions

Technological advancement is increasingly reliant on scientific discoveries, yet the mechanisms driving this reliance remain unclear. Grounded in knowledge recombination theory, this study utilizes a dataset of 1,685,970 invention patents granted by the USPTO from 2001 to 2010. Through regression models including Logit and Tobit, this study examines the effects of knowledge recombination and invention team characteristics on the propensity and intensity of scientific knowledge incorporation. This study yields several interesting findings.

Firstly, this study reveals that the complexity and novelty of recombination significantly promote both the tendency and intensity of scientific knowledge integration. Specifically, the higher the complexity and novelty of a technological invention's recombination, the more likely it is to utilize scientific knowledge in the invention process, and the greater the number of scientific knowledge components involved. As previously mentioned, it is evident that more complex inventions engage a wider variety of technological knowledge components throughout the knowledge recombination phase. Fleming and Sorenson (2004) noted that scientific knowledge enables inventors to overcome the inherent difficulties associated with integrating highly coupled technological components. Consequently, the greater the complexity of a technology, the more likely it is that inventors will incorporate scientific knowledge components during the technological development process. Similarly, the higher the novelty of knowledge recombination in a technology, the more it indicates the integration of technologically distant knowledge components during its invention process. This increased distance necessitates a greater input of scientific knowledge to mitigate the uncertainties and identification challenges that such distant recombination might introduce (Balconi et al., 2010).

Secondly, the knowledge maturity exists a significant negative impact on both the tendency and intensity of scientific knowledge integration. This suggests that more mature technologies are less likely to incorporate scientific knowledge. As demonstrated by Chin et al. (2012), in more mature technologies, domain-specific literature increasingly becomes common knowledge. Inventors often internalize this information as part of their foundational skills, consequently reducing their reliance on citing scientific literature during the patent application process. Therefore, as technology matures, its dependency on scientific knowledge diminishes.

Thirdly, the science-based invention experience of invention teams demonstrates a significant positive impact on both the tendency and intensity of scientific knowledge integration. As evidenced by the marginal effects presented in Table 3, the science-based invention experience of the inventor team emerges as the most critical factor influencing the propensity for scientific knowledge input. Specifically, each unit increase in the inventors' science-based invention experience is associated with a 32.6% increase in the probability of leveraging scientific knowledge in technological inventions. This phenomenon may be attributed to the inventors' prior successful experiences with science-based inventions, which potentially enhance their ability to identify connections between past experiences and current challenges, as well as to conceptualize potential solutions. In contrast, the breadth of inventor teams' technological knowledge exhibits a significant negative impact on both the tendency and intensity of scientific knowledge integration. That is, the broader the range of technological knowledge possessed by the invention team, the

less likely they are to utilize scientific knowledge in addressing specific technical challenges. The individual knowledge base serves as the fundamental wellspring of innovation (Boh et al., 2014; Castaneda and Cuellar, 2020; Zhu et al., 2023). Inventors typically initiate their problem-solving process by exploring their existing knowledge domains, seeking to identify and synthesize knowledge components that address specific technological challenges. The broader the technological knowledge base of an invention team, the more diverse knowledge combination strategies they can employ, consequently reducing their reliance on scientific knowledge components.

In contrast to many existing studies, our primary focus is the nature of science-based technologies, rather than the subsequent impacts of inventions. Inventive activities involve various interconnected stages, ranging from conceptualization to market development. Knowledge recombination characteristics of inventions can potentially influence the entire innovation chain. Technological complexity, novelty, and maturity have been shown in prior studies to explain their knowledge spill over effect on the value of subsequent inventions to varying degrees (Barbieri at al., 2020; Capaldo et al., 2017). Therefore, re-examining previous empirical research by taking these factors into account can shed new light on the impact of science on technological inventions, which also serves as a direction for our future research.

Our study is not without limitations. We solely rely on the USPTO patent data for the empirical analyses. In fact, there exist disparities in the patent systems of different countries. USPTO has been identified as exhibiting a greater frequency of citations than patents from other patent offices (Svensson, 2022) and US generated about 63% of the world's science-based patents (Gazni and Ghaseminik, 2019). Future research could investigate the robustness of our findings across different patent offices. Likewise, the degree of science dependency within various technological fields should also be considered. Technological inventions with high scientific dependence, such as those in biomedical and chemical fields, may present distinct recombination patterns than patents in other fields. Additionally, while our study examines the individual effects of knowledge recombination characteristics and inventor team attributes, future research could explore their interactive effects to better understand how they jointly shape scientific knowledge integration in invention processes.

Acknowledgement

This study is support by the National Natural Science Foundation of China (72232006); The Natural Science Foundation of Hubei province (Grant No.2024AFB393); the National Natural Science Foundation of China (72204189).

About the authors

Guiyan Ou is a post-doc researcher in the School of Information Management, Wuhan University, Wuhan, China. She received her Ph.D. from Wuhan University, and her research interests include scientometrics and patent innovation. She can be contacted at ouguiyan@whu.edu.cn

Haodong Chen is a PhD student in the School of Information Management, Wuhan University, Wuhan, China. His research interests are in the digital transformation and innovation management. He can be contacted at ghauxtungc@whu.edu.cn.

Kaili Wang is a Lecturer in the Faculty of Artificial Intelligence in Education of Central China Normal University Wuhan, China. She received her Ph.D. from Wuhan University, and her research interests are in the Policy Informatics, Information Systems. She can be contacted at kailiw@ccnu.edu.cn

Jiang Wu is the Professor in the School of Information Management, Wuhan University, Wuhan, China. He received their Ph.D. from the Huazhong University of Science and Technology, and their research interests are in the social network computing; network information measurement, He can be contacted at jiangw@whu.edu.cn

References

Afuah, A., & Tucci, C. L. (2012). Crowdsourcing as a solution to distant search. Academy of management review, 37(3), 355-375. https://doi.org/10.5465/amr.2010.0146

Ahuja, G., & Morris Lampert, C. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. Strategic management journal, 22(6-7), 521-543. https://doi.org/10.1002/smj.176

Aldrich, H. E., & Fiol, C. M. (1994). Fools rush in? The institutional context of industry creation. Academy of management review, 19(4), 645-670. https://doi.org/10.5465/amr.1994.9412190214

Anckaert, P. E., Cassiman, D., & Cassiman, B. (2020). Fostering practice-oriented and use-inspired science in biomedical research. Research Policy, 49(2), 103900. https://doi.org/10.1016/j.respol.2019.103900

Ardito, L., Petruzzelli, A. M., & Albino, V. (2016). Investigating the antecedents of general-purpose technologies: A patent perspective in the green energy field. Journal of Engineering and Technology Management, 39, 81-100. https://doi.org/10.1016/j.jengtecman.2016.02.002

Arora, A., Belenzon, S., & Dionisi, B. (2023). First-mover advantage and the private value of public science. Research Policy, 52(9), 104867. https://doi.org/10.1016/j.respol.2023.104867

Arora, A., & Gambardella, A. (1994). The changing technology of technological change: general and abstract knowledge and the division of innovative labour. Research policy, 23(5), 523-532. https://doi.org/10.1016/0048-7333(94)01003-X

Balconi, M., Brusoni, S., & Orsenigo, L. (2010). In defence of the linear model: An essay. Research policy, 39(1), 1-13. https://doi.org/10.1016/j.respol.2009.09.013

Barbieri, N., Marzucchi, A., & Rizzo, U. (2020). Knowledge sources and impacts on subsequent inventions: Do green technologies differ from non-green ones?. Research Policy, 49(2), 103901. https://doi.org/10.1016/j.respol.2019.103901

Baumann, O., & Siggelkow, N. (2013). Dealing with complexity: Integrated vs. chunky search processes. Organization Science, 24(1), 116-132. https://doi.org/10.1287/orsc.1110.0729

Boh, W. F., Evaristo, R., & Ouderkirk, A. (2014). Balancing breadth and depth of expertise for innovation: A 3M story. Research Policy, 43(2), 349-366. https://doi.org/10.1016/j.respol.2013.10.009

Bozeman, B. (2000). Technology transfer and public policy: a review of research and theory. Research policy, 29(4-5), 627-655. https://doi.org/10.1016/S0048-7333(99)00093-1

Capaldo, A., Lavie, D., & Messeni Petruzzelli, A. (2017). Knowledge maturity and the scientific value of innovations: The roles of knowledge distance and adoption. Journal of Management, 43(2), 503-533. https://doi.org/10.1177/0149206314535

Castaneda, D. I., & Cuellar, S. (2020). Knowledge sharing and innovation: A systematic review. Knowledge and Process Management, 27(3), 159-173. https://doi.org/10.1002/kpm.1637

Chen, T., Kim, C., & Miceli, K. A. (2021). The emergence of new knowledge: The case of zero-reference patents. Strategic Entrepreneurship Journal, 15(1), 49-72. https://doi.org/10.1002/sej.1385

Chin, Y. L., Wu, F. S., Lin, T. C., Lin, B. W., & Chan, T. Y. (2012, July). Why patents have lower citation on non-patent references?: A case study from Taiwan. In 2012 Proceedings of PICMET'12: Technology Management for Emerging Technologies (pp. 1054-1059). IEEE.

Choi, H., Shin, J., & Hwang, W. S. (2018). Two faces of scientific knowledge in the external technology search process. Technological Forecasting and Social Change, 133, 41-50. https://doi.org/10.1016/j.techfore.2018.02.020

Du Plessis, M., Van Looy, B., Song, X., & Magerman, T. (2010). Data production methods for harmonized patent statistics: Patentee sector allocation 2009. Eurostat Working Paper-Annual Report 2009.

Dunne, D. D., & Dougherty, D. (2016). Abductive reasoning: How innovators navigate in the labyrinth of complex product innovation. Organization Studies, 37(2), 131-159. https://doi.org/10.1177/0170840615604450

Fleming, L., Mingo, S., & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. Administrative science quarterly, 52(3), 443-475. https://doi.org/10.2189/asqu.52.3.443

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. Strategic management journal, 25(8-9), 909-928. https://doi.org/10.1177/01492063211055982

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. Research policy, 30(7), 1019-1039. https://doi.org/10.1016/S0048-7333(00)00135-9

Fusillo, F. (2023). Green Technologies and diversity in the knowledge search and output phases: Evidence from European Patents. Research Policy, 52(4), 104727. https://doi.org/10.1016/j.respol.2023.104727

Gazni, A., & Ghaseminik, Z. (2019). The increasing dominance of science in the economy: Which nations are successful?. Scientometrics, 120(3), 1411-1426. https://doi.org/10.1007/s11192-019-03161-5

Gruber, M., Harhoff, D., & Hoisl, K. (2013). Knowledge recombination across technological boundaries: Scientists vs. engineers. Management Science, 59(4), 837-851. https://doi.org/10.1287/mnsc.1120.1572

Grupp, H., & Schmoch, U. (1992). Perceptions of scientification of innovation as measured by referencing between patents and papers: dynamics in science-based fields of technology. In Dynamics of science-based innovation (pp. 73-128). Berlin, Heidelberg: Springer Berlin Heidelberg.

Guo, B., Gao, J., & Chen, X. (2013). Technology strategy, technological context and technological catch-up in emerging economies: industry-level findings from Chinese manufacturing. Technology analysis & strategic management, 25(2), 219-234. https://doi.org/10.1080/09537325.2012.759201

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition, and the value of patent rights. Research policy, 32(8), 1343-1363. https://doi.org/10.1016/S0048-7333(02)00124-5

Heeley, M. B., & Jacobson, R. (2008). The recency of technological inputs and financial performance. Strategic Management Journal, 29(7), 723-744. https://doi.org/10.1002/smj.682

Hohberger, J. (2016). Diffusion of science-based inventions. Technological Forecasting and Social Change, 104, 66-77. https://doi.org/10.1016/j.techfore.2015.11.019

Hur, W., & Oh, J. (2021). A man is known by the company he keeps?: A structural relationship between backward citation and forward citation of patents. Research Policy, 50(1), 104117. https://doi.org/10.1016/j.respol.2020.104117

Jiao, H., Wang, T., & Yang, J. (2022). Team structure and invention impact under high knowledge diversity: An empirical examination of computer workstation industry. Technovation, 114, 102449. https://doi.org/10.1016/j.technovation.2021.102449

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. Strategic Management Journal, 36(10), 1435-1457. https://doi.org/10.1002/smj.2294

Katila, R., & Ahuja, G. (2002). Something old, something new: A longitudinal study of search behavior and new product introduction. Academy of management journal, 45(6), 1183-1194. https://doi.org/10.1002/smj.4250171004

Kauffman, S. A. (1993). The origins of order: Self-organization and selection in evolution. Oxford University Press.

Keijl, S., Gilsing, V. A., Knoben, J., & Duysters, G. (2016). The two faces of inventions: The relationship between recombination and impact in pharmaceutical biotechnology. Research Policy, 45(5), 1061-1074. https://doi.org/10.1016/j.respol.2016.02.008

Lee, K., & Lim, C. (2001). Technological regimes, catching-up and leapfrogging: findings from the Korean industries. Research policy, 30(3), 459-483. https://doi.org/10.1016/S0048-7333(00)00088-3

Levinthal, D. A., & March, J. G. (1993). The myopia of learning. Strategic management journal, 14(S2), 95-112. https://doi.org/10.1002/smj.4250141009

Liao, S. H., Fei, W. C., & Liu, C. T. (2008). Relationships between knowledge inertia, organizational learning and organization innovation. Technovation, 28(4), 183-195. https://doi.org/10.1016/j.technovation.2007.11.005

Maddala, G. S. (1987). Limited dependent variable models using panel data. Journal of Human resources, 307-338. https://doi.org/10.2307/145742

March, J. G. (1991). Exploration and exploitation in organizational learning. Organization science, 2(1), 71-87. https://doi.org/10.1287/orsc.2.1.71

Marx, M., & Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. Strategic Management Journal, 41(9), 1572-1594. https://doi.org/10.1002/smj.3145

Meyer, M. (2000). Does science push technology? Patents citing scientific literature. Research policy, 29(3), 409-434. https://doi.org/10.1016/S0048-7333(99)00040-2

Nightingale, P. (1998). A cognitive model of innovation. Research policy, 27(7), 689-709. https://doi.org/10.1016/S0048-7333(98)00078-X

Nooteboom, B. (2000). Learning and innovation in organizations and economies. OUP Oxford.

Novelli, E. (2015). An examination of the antecedents and implications of patent scope. Research Policy, 44(2), 493-507. https://doi.org/10.1016/j.respol.2014.09.005

Perry-Smith, J. E., & Shalley, C. E. (2014). A social composition view of team creativity: The role of member nationality-heterogeneous ties outside of the team. Organization Science, 25(5), 1434-1452. https://doi.org/10.1287/orsc.2014.0912

Petruzzelli, A. M., Ardito, L., & Savino, T. (2018). Maturity of knowledge inputs and innovation value: The moderating effect of firm age and size. Journal of Business Research, 86, 190-201. https://doi.org/10.1016/j.jbusres.2018.02.009

Petruzzelli, A. M., & Savino, T. (2014). Search, recombination, and innovation: Lessons from haute cuisine. Long Range Planning, 47(4), 224-238. https://doi.org/10.1016/j.lrp.2012.09.001

Poege, F., Harhoff, D., Gaessler, F., & Baruffaldi, S. (2019). Science quality and the value of inventions. Science advances, 5(12), eaay7323. DOI: 10.1126/sciadv.aay7323

Rosenkopf, L., & Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. Strategic management journal, 22(4), 287-306. https://doi.org/10.1002/smj.160

Savino, T., Messeni Petruzzelli, A., & Albino, V. (2017). Search and recombination process to innovate: a review of the empirical evidence and a research agenda. International Journal of Management Reviews, 19(1), 54-75. https://doi.org/10.1111/ijmr.12081

Saviotti, P. P. (1996). Technological evolution, variety, and the economy. In Technological Evolution, Variety, and the Economy. Edward Elgar Publishing.

Simonton, D. K. (2003). Scientific creativity as constrained stochastic behavior: the integration of product, person, and process perspectives. Psychological bulletin, 129(4), 475. DOI: 10.1037/0033-2909.129.4.475

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). Measuring patent quality: Indicators of technological and economic value. https://doi.org/10.1787/18151965

Sung, H. Y., Wang, C. C., Huang, M. H., & Chen, D. Z. (2015). Measuring science-based science linkage and non-science-based linkage of patents through non-patent references. Journal of Informetrics, 9(3), 488-498. https://doi.org/10.1016/j.joi.2015.04.004

Svensson, R. (2022). Patent value indicators and technological innovation. Empirical Economics, 62(4), 1715-1742. https://doi.org/10.1007/s00181-021-02082-8

Sørensen, J. B., & Stuart, T. E. (2000). Aging, obsolescence, and organizational innovation. Administrative science quarterly, 45(1), 81-112. https://doi.org/10.2307/2666980

Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. Economics of Innovation and new technology, 5(1), 19-50. https://doi.org/10.1080/10438599700000006

Turner, S. F., Mitchell, W., & Bettis, R. A. (2013). Strategic momentum: How experience shapes temporal consistency of ongoing innovation. Journal of Management, 39(7), 1855-1890. https://doi.org/10.1177/014920631245870

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. Research Policy, 46(8), 1416-1436.https://doi.org/10.1016/j.respol.2017.06.006

Xiao, T., Makhija, M., & Karim, S. (2022). A knowledge recombination perspective of innovation: review and new research directions. Journal of Management, 48(6), 1724-1777. https://doi.org/10.1177/01492063211055982

Zhao, Q., Luo, Q., & Tao, Y. (2023). The power of paper: Scientific disclosure and firm innovation. Finance Research Letters, 56, 104147. https://doi.org/10.1016/j.frl.2023.104147

Zhu, J. X., Sun, M., Wei, S. X., & Ye, F. Y. (2023). Characterizing patent big data upon IPC: a survey of triadic patent families and PCT applications. Journal of Big Data, 10(1), 85. https://doi.org/10.1186/s40537-023-00778-5

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Are we there yet? Evaluation of AI-generated metadata for online information resources

Vyacheslav Zavalin and Oksana L. Zavalina

DOI: https://doi.org/10.47989/ir30iConf47215

Abstract

Introduction. Generative AI tools are increasingly used in creating descriptive metadata the quality of which is key for information discovery and support of information user tasks. Machine-readable online information resources such as websites naturally lend themselves to automatic metadata creation. Yet, assessments of AI-generated metadata for them are lacking. AI metadata quality research to date is limited to 2 metadata standards.

Method. This experimental study assessed the quality of AI-generated descriptive metadata in 4 most widely used standards: Dublin core, MODS, MARC, and BIBFRAME. Three generative AI tools – Gemini, Gemini advanced, and ChatGPT4 – were used to create metadata for an educational website.

Analysis. Zero-shot queries prompting AI tools to generate metadata followed the same structure and included the link to metadata scheme's openly accessible documentation. Comparative in-depth analysis of accuracy and completeness of entire resulting AI-generated metadata records was performed.

Results. Overall, AI-generated metadata does not meet the quality threshold. ChatGPT performs somewhat better than 2 other tools on completeness, but accuracy is similarly low in all 3 tools.

Conclusions. Current metadata-generating effectiveness of AI tools does not allow to conclude that involvement of human metadata experts in creation of quality (and therefore functional) metadata can be significantly reduced without strong negative impact on information discovery.

Introduction

Artificial intelligence (AI) generative tools are increasingly experimented with to assess AI usability in assisting humans in descriptive metadata creation. Metadata is crucial for discoverability of all kinds of information resources. Major metadata standards designed with the goal of describing (with relatively small-scale sets of metadata elements) and providing access to online information resources include 2 versions of Dublin Core (simple (https://www.dublincore.org/specifications/dublin-core/dces/) and extended (https://www.dublincore.org/specifications/dublin-core/dcmi-terms)), and metadata object description schema (MODS) (https://loc.gov/standards/mods/) introduced in late 1990s and evolving since then. In addition, the key metadata standard in use by libraries worldwide (with hundreds of millions of metadata records currently aggregated in WorldCat database (https://search.worldcat.org/)) since late 1970s, machine readable cataloguing (MARC) (https://www.loc.gov/marc/bibliographic/), originally intended for representing analogue (text and audio-visual) information resources, is a much more comprehensive metadata element set that has been continuously evolving into a standard that supports discovery of any type of information resources, including those available online in any file format (e.g., HTML, streaming audio and video, PDF, etc.). BIBFRAME (https://www.loc.gov/bibframe/) – launched in 2010s and intended as an alternative to MARC – is both an ontology and a rich metadata element set designed for enabling seamless resource discovery by emphasizing relationships representation and use of Linked Data approaches in describing information resources and entities related to them: their creators, contributors, subjects, etc. Regardless of the metadata element set used, metadata creators consult guidelines designed iteratively by teams of experts and approved by metadata community of practice to support decision-making by those tasked with describing information resources to facilitate discovery and access.

Literature Review

Metadata's effectiveness in supporting information resource discovery, as well as specific information user tasks formulated in the IFLA library reference model (LRM) is the defining feature of quality metadata (Riva, Lebeouf, & Žumer, 2017). Metadata quality evaluations' results inform improvements in metadata education and practice which translate into higher quality, more functional metadata. Evaluations are guided by metadata quality frameworks that define quality criteria and propose measures for assessing metadata against these criteria. The most influential framework developed by Bruce and Hillman (2004) consists of 7 criteria. Over the years, these criteria have been adopted and adapted for analysis in many metadata studies, and informed development of other frameworks, most notably the information quality framework by Stvilia and colleagues (2007), which consists of 22 criteria. Three metadata quality criteria defined in these frameworks – *accuracy*, *completeness*, and *consistency* – are used in studies the most often, and according to the survey of metadata managers (Park & Tosaka, 2010) are considered the most important in quality assurance.

Most metadata evaluations so far assessed human-created metadata because the lion's share of metadata that supports discovery is created this way (even when automation tools are used to assist in this job, and even despite the recent advances of AI generative tools that offer a potential of partially automating the process) and will likely remain largely human-generated. It is important to remember that ideas behind using machines to automate the complex tasks of metadata creation are not new and certainly not unique to the current AI revolution. Decades before the generative AI emergence, machine-generated metadata was widely researched, developed, tested, and implemented (as shown, for example, in overview of automatic metadata generation projects in Greenberg, 2005). For instance, various tools designed for automating some of the relatively small-scale (in terms of the space, measured in number of characters occupied in the metadata record) but important and intellectually challenging metadata creation tasks (e.g., generating a

classification number based on the included subject headings, etc.) have been in existence for decades and used with acceptable accuracy – by many professional metadata creators to reduce the amount of time needed for metadata record creation (e.g., Golub, 2021). Other popular uses of such tools were to automatically add to the metadata records certain codes (e.g., those representing languages used in the information resource and geographic areas covered by the content of the resource). These simple tools often take the form of macros that are shared in the professional community of metadata creators and are applied to represent any kind of information resource using the subscription based OCLC Connexion professional software (e.g., OCLC, 2023).

Other electronic tools designed to assist metadata creation are openly available to anyone interested and take the shape of online forms, to be completed, based on guidelines, by a human. The information entered by trained metadata creators into such form is then processed and converted into a certain encoding (e.g., JSON-LD, RDFXML, etc.). The most widely known such tool is the BIBFRAME Editor (https://bibframe.org/marva/editor/) designed to support human expert work with BIBFRAME metadata element set and guidelines. This tool has several application profiles built into it for representing monographs, notated music, serials, cartographic resources, sound recordings, moving images, rare materials, prints and photographs. Another notable openly accessible tool that is a set of online forms (for representing datasets, e-books, government documents, maps, microfilms, monographs, scores, serials, and theses and dissertations) is Metadata Maker (https://metadatamaker.library.illinois.edu/) developed by the University of Illinois at Urbana-Champaign Library based on BIBFRAME 2.0 (Heng & Han, 2023; Michael & Han, 2020). This tool allows creating metadata in MARC, MARCXML, MODS, and ONIX metadata schemes.

Due to their very nature, machine-readability, and internet indexability, online information resources such as webpages are presumably much easier to automate the metadata creation processes for and substantially reduce the amount of human effort needed. With that goal, long before the emergence of AI generative tools, the DC-dot (https://www.ukoln.ac.uk/metadata/dcdot/) automation tool was developed – using Java Script – and made available in the 1990s by the Innovation Support Centre at The United Kingdom Office for Library and Information Networking (UKOLN). The DC-dot tool generated metadata records encoded in either HTML or RDFXML and was widely used by metadata creators for generating an entire metadata record (albeit basic, and in need of verification and augmentation by human experts) in simple version of Dublin Core standard that includes 15 metadata elements. Some versions of this popular tool were integrated in professional workflow in OCLC Connexion – the metadata management utility that feeds WorldCat database. However, due to '*the cessation of core funding*' in 2013, this popular tool that consistently generated reliable metadata results for webpages was discontinued. Ten years later, the AI generative tools became available, with the promise of offering not only the functionality of DC-dot but more robust automatic metadata generation use, capable of creating metadata records for various kinds of information resources and following various metadata schemes.

Most published studies which assess descriptive metadata quality focus on Dublin Core, MODS, or MARC (e.g., Jackson et al., 2008; Kurtz, 2010; Park & Maszaros, 2009; Weagley, Gelches & Park, 2010; Zavalin, Zavalina & Safa, 2021; Zavalin, Zavalina & Miksa, 2021). Likely due to the much younger age of BIBFRAME metadata element set and the vast complexity and length of BIBFRAME metadata records – which, when of a reasonable quality, are at least three times longer than MARC records for the same information resource represented in XML syntax (https://id.loc.gov/tools/bibframe/comparebf-lccn/2018958785.xml,) - no published studies have examined the quality of BIBFRAME metadata.

AI's potential uses in metadata generation have been explored by multiple authors who focused on individual specific metadata tasks such as automatic assignment of subject headings (e.g., Chou &

Chu, 2022; Chow, Kao, & Li, 2024; Ganadi et al., 2023) or classification numbers (Bodenhamer, 2023; Martorana, et al., 2024; Zhang, Wu, & Zhang, 2023). Two studies evaluating AI-generated descriptive metadata records have been published at the time of writing this paper, to the end of our knowledge. Taniguchi (2024) compared ChatGPT-created and human-expert-created MARC records for maps, sheet music, sound recordings, etc. Brzustowitz (2023) used in-depth comparative analysis of small samples of MARC and Dublin Core records created by human metadata experts and ChatGPT to represent books and audio recordings. As of the time of writing our paper, no studies have explored the quality of AI-generated metadata representing online resources and no study looked beyond MARC and simple Dublin Core in AI-generated analysis. Our project preliminary results of which are reported below addresses this research gap.

Method

In this study, we used basic and advanced versions of Google's Gemini AI tool, as well as ChatGPT 4, to generate AI metadata for the same resource (an English-language educational webpage (https://informationscience.unt.edu/)) in 4 most widely used metadata schemes / element sets (in the order of simplest-to-most-complex):

•Extended version of Dublin Core: Dublin Core Metadata Initiative (DCMI) Metadata Terms

•Metadata Object Description Schema (MODS)

•MARC 21 Bibliographic Format, and

•BIBFRAME.

Google's two AI generative tools – Gemini and Gemini Advanced – were selected for our study, in addition to ChatGPT that is the most used AI-generative tool in existing research on AI usability in metadata generation. The reason for Gemini and Gemini Advanced selection was that this study's target information resource type is a website indexed by Google. Thus, the assumption was that these Google AI tools can create good quality metadata by harnessing the power of Google-indexed webpages.

We analysed the quality of the results, relying on our team's expert knowledge of these metadata elements sets and existing guidelines and documentation for these standards, as well as on expert-human-created metadata for the target information resource used in this experiment, with the goal of answering the following research questions:

•How *accurate* is the AI-generated metadata?

•How *complete* is the AI-generated metadata?

•How does the quality of AI-generated metadata compare across the 4 metadata element sets and 3 AI generative tools?

We evaluated completeness and accuracy based on the measures detailed in the recent studies analysing quality of metadata in Dublin Core and other metadata schemes (e.g., Aljalahmah & Zavalina, 2024; Zavalin & Zavalina, 2023; Zavalina & Burke, 2021). The measures included:

- for *completeness*: the total number of metadata fields and their instances in the record, the number of missing applicable metadata fields/instances, etc.
- for *accuracy*: the use of appropriate encoding, well-formedness, and validity of XML; the number of metadata fields and instances containing errors: misrepresentation of information resource in the date value, using non-authorized term from a controlled vocabulary, misusing the field that is intended for other kinds of data, etc.

The official documentation for the examined metadata standards and human-created metadata records for the target information resource served as a source of information on the number of applicable metadata fields, etc. that were used in analysis. The prompts used in all 3 generative AI tools followed this pattern: '*Generate a metadata record for this webpage: https://informationscience.unt.edu/. Follow the [NAME & VERSION] standard: [LINK TO STANDARD]*'. For all 4 metadata schemes, the official documentation is freely available online in machine-readable HTML or XML format.

Preliminary Results and Discussion

Figure 1 presents the prompts that were used in this experiment and summarizes the results of the quality assessment of the outputs generated by Gemini, Gemini Advanced, and ChatGPT. Overall, metadata records generated by ChatGPT were more detailed. Those generated by Gemini Advanced for 3 metadata schemes – Dublin Core, MODS, and MARC – tended to contain more details than the ones generated by Gemini. This is in line with our assumptions based on Google's positioning Gemini Advanced as a more powerful AI tool.

Dublin Core metadata generated by ChatGPT and Gemini Advanced, while incomplete, was comparable in the number of metadata fields included to OAIster-harvested Dublin Core records that previous research found to have 8 metadata fields in most cases (Jackson et al., 2008). However, it is important to note that OAIster aggregation metadata follows the simple version of Dublin Core, while this experiment tasked AI generative tools with creating metadata that follows the extended version of Dublin Core standard (DCMI Metadata Terms) which has over 40 metadata elements as opposed to 15. For more robust MODS and MARC metadata schemes, completeness was moderate (with a few important omissions) in ChatGPT-generated record, and very low in both versions of the Gemini tool, although somewhat more complete in Gemini Advanced (Figure 1).

While ChatGPT and Gemini Advanced generated BIBFRAME records (albeit incomplete, and with basic JSON encoding as opposed to JSON-LD or RDF serialization that support Linked Data functionality), Gemini returned this statement,

> *I don't have a tool that can directly generate bibframe metadata for webpages, but I can search on the web using google tools. Here's what I found on the web: I couldn't find the bibframe metadata for this webpage, but there are some resources that discuss bibframe [...].*

The accuracy was low in all 3 generative-AI tools' outputs for Dublin Core, MODS, and MARC. The proportion of metadata fields with mistakes was similarly high across tools for the same metadata scheme. Also, all 3 tools failed to generate the DCMI Metadata Terms records and generated records that follow the simpler version of this same metadata scheme: DCMES 1.1. However, unexpectedly, for Dublin Core and MODS, the simplest of the 3 tools, Gemini was more successful than Gemini Advanced in generating the metadata record properly encoded (in XML) and structured (with both top-level elements and sub elements) for MODS. ChatGPT was successful in structuring the metadata record in this hierarchical metadata scheme. ChatGPT-generated metadata was encoded in machine-readable syntax for all 4 metadata schemes (XML for Dublin Core and MODS, MARC for MARC Bibliographic Format, and JSON for BIBFRAME). However, XML encoding was missing key components that negatively affected its functionality in information discovery and wrong version of JSON syntax was used (Figure 1).

Metadata Scheme	Summary of AI-generated metadata quality: Gemini	Summary of AI-generated metadata quality: Gemini Advanced	Summary of AI-generated metadata quality: ChatGPT 4
Dublin Core DCMI Metadata Terms	**Completeness**: 5 metadata fields (5 field instances total used out of 19 applicable fields **XML encoding**: partially present but missing key parts of XML document (XML declaration and prologue) **Accuracy**: incorrect version of Dublin Core (DCMES 1.1) used; mistakes in 3 fields	**Completeness**: 9 metadata fields (9 instances) used out of 19 applicable fields **XML encoding:** absent (non-machine-readable record in table form) **Accuracy**: incorrect version of Dublin Core (DCMES 1.1) used; mistakes in 2 fields	**Completeness**: 13 metadata fields (13 instances) used out of 19 applicable fields **XML encoding:** partially present but missing key parts of XML document (XML declaration and prologue) **Accuracy**: incorrect version of Dublin Core (DCMES 1.1) used; mistakes in 3 fields
Metadata Object Description Schema (MODS) 3.8	**Completeness**: 4 top-level metadata fields (4 instances; with some subfields) used out of 15 applicable fields **XML encoding:** present **Accuracy**: mistakes in 3 fields	**Completeness**: 8 top-level metadata fields (8 instances; missing ALL subfields) used out of 15 applicable fields **XML encoding:** absent (non-machine-readable record in table form) **Accuracy**: mistakes in 4 fields	**Completeness**: 12 top-level metadata fields (12 instances; with most subfields) used out of 15 applicable fields **XML encoding:** present **Accuracy**: mistakes in 5 fields
MARC 21 Bibliographic Format	**Completeness**: 2 metadata fields (2 instances) used out of 24 applicable fields. **MARC encoding:** present **Accuracy**: mistakes in 1 field	**Completeness**: only 9 metadata fields (16 instances) out of 24 applicable fields **MARC encoding:** present **Accuracy**: mistakes in 4 fields (10 instances)	**Completeness**: only 9 metadata fields (9 instances) out of 24 applicable fields **MARC encoding:** present **Accuracy**: mistakes in 4 fields (4 instances)
BIBFRAME metadata element set	***Failure to generate metadata***	**Completeness:** 5 metadata fields **BIBFRAME encoding:** absent **Accuracy**: record does NOT follow BIBFRAME and is a lower-quality version of Dublin Core record	**Completeness:** only 6 metadata fields (9 instances) - less complete than MARC, MODS, or Dublin Core record **BIBFRAME encoding:** present **Accuracy:** uses simple JSON syntax (not any RDF serialization – e.g., RDFXML– or JSON-LD which support Linked Data functionality)

Figure 1. Metadata quality evaluation results summary

Conclusions and next steps

The overall quality of AI-generated metadata in this experiment does NOT meet the basic expectations for functional metadata. Low-quality metadata – such as observed for AI-generated metadata in this study - fails to support Find, Identify, Select, Obtain, & Explore user tasks defined in IFLA LRM (Riva, Lebeouf, & Žumer, 2017). At this point in the development of AI tools, relying solely on AI-generated metadata creation would result in significantly compromised information access. However, AI-generated metadata could be used as a starting point in the human metadata creation for Dublin Core, MARC, and MODS metadata standards, with the need for thorough checking and editing by human metadata creators. AI-generated BIBFRAME metadata with Gemini

and Gemini Advanced Google AI tools is currently non-existent, while ChatGPT-produced BIBFRAME metadata although very incomplete and not supporting Linked Data functionality, could be used by expert metadata professionals as a starting point, with significant manual revisions.

This exploratory research focused on one type of information resource previously excluded from experimenting with AI metadata generation: online educational website. Future analysis will compare accuracy and completeness of metadata generated by these same AI tools for different types of information resources. Another important direction for further research is training the AI-generative tools by refining the metadata generation prompts in different ways and by providing the tool (where applicable) with feedback on the quality of generated metadata.

Findings of studies such as this one, as well as experimenting with AI-generated metadata itself, need to be integrated in metadata education for information professionals. For example, instructors of cataloguing and digital library metadata courses can demonstrate to their students how these tools work for metadata generation, use AI-generated metadata in teaching students the topics of metadata quality evaluation, ask students to improve AI-generated records as part of their metadata-creation assignments.

The need for research into the quality of AI-generated descriptive metadata is time-sensitive because '*AI has been increasingly shaping the library management landscape in recent* years' (Bisht et al., 2023). While we agree that '*integration of AI technology has the potential to greatly enhance the efficiency, accuracy, and user experience in library cataloguing*' (Bisht et al., 2023) by assisting the human metadata creators, this potential needs to be thoroughly and objectively evaluated before taking any related management steps. It is important to conduct such examinations to collect the data that should be used in evidence-based decisions regarding management of information organizations (including libraries) and such crucial functions as metadata creation.

About the authors

Vyacheslav Zavalin is Assistant Professor in Department of Information Science, Texas Woman's University of North Texas, USA. He received his Ph.D. from University of North Texas, and his research interests are in cataloguing, metadata, and data analytics. Dr. Zavalin can be contacted at vzavalin@twu.edu

Oksana L. Zavalina is Professor in Department of Information Science, University of North Texas, USA. She received her Ph.D. from University of Illinois, and her research interests are in the information organization in libraries and other repositories, assessment of metadata quality and user needs. Dr. Zavalina can be contacted at Oksana.Zavalina@unt.edu

References

Aljalahmah, S.H., & Zavalina, O. L. (2024). Student-created Dublin Core metadata representing Arabic language eBooks: Comparison of individual and group work outcomes. Journal of Education for Library and Information Science (JELIS), 65 (3), 325-344. https://doi.org/10.3138/jelis-2023-0016

Bisht, S., Nutiyal, A.P., Sharma, S., Sai, M., Bathla, N., & Singh, P. (2023). The role of Artificial Intelligence in shaping library management and its utilization. In, 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India (pp. 467-472). https://doi.org/10.1109/ICDT57929.2023.10150520.

Bodenhamer, J. (2023). The reliability and usability of ChatGPT for library metadata. https://hdl.handle.net/11244/339626 (accessed 12 August 2024)

Bruce, T.R., & Hillman, D.I. (2004). The continuum of metadata quality: defining, expressing, exploiting. In: Hillman, D., Westbrook, L. (eds.) Metadata in Practice, pp. 238–256. American Library Association, Chicago.

Brzustowicz, R. (2023). From ChatGPT to CatGPT: The implications of Artificial Intelligence on library cataloguing. Information Technology and Libraries, 42(3). https://doi.org/10.5860/ital.v42i3.16295

Chou, C., & Chu, T. (2022). An analysis of BERT (NLP) for assisted subject indexing for project Gutenberg. Cataloging & Classification Quarterly, 68 (8), 807-835. https://doi.org/10.1080/01639374.2022.2138666

Chow, E. H. C., Kao, T. J., & Li, X. (2024). An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations. Cataloging & Classification Quarterly, 62(5), 574–588. https://doi.org/10.1080/01639374.2024.2394516

Golub, K. (2021). Automated subject indexing: An overview. Cataloging & Classification Quarterly, 59(8), 702–719. https://doi.org/10.1080/01639374.2021.2012311

Greenberg, J. (2005). Metadata generation: Processes, people, and tools. Bulletin of the American Society for Information Science and Technology. https://doi.org/10.1002/bult.269

Heng, G., & Han, M.J. (2023). Revamping Metadata Maker for 'Linked Data Editor': Thinking Out Loud. Code4Lib Journal, 55. https://journal.code4lib.org/articles/16925

Ganadi, A.E., Vigliermo, R.A., Sala, L., Vanzini, M., Ruozzi, F., & Bergamaschi, S. (2023). Bridging Islamic knowledge and AI: inquiring ChatGPT on possible categorizations for an Islamic digital library. Proceedings of the 2nd Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2023) co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), November 6, 2023, Roma, Italy. https://ceurws.org/Vol-3536/03_paper.pdf (accessed 12 August 2024)

Jackson, A.S., Han, M.J., Groetsch, K., Mustafoff, M., & Cole, T.W. (2008). Dublin Core metadata harvested through OAI-PMH. Journal of Library Metadata, 8(1), 5–21. https://doi.org/10.1300/J517v08n01_02

Kurtz, M. (2010). Dublin Core, DSpace, and a brief analysis of three university repositories. Information Technology and Libraries, 29(1), 40–46). https://doi.org/10.6017/ital.v29i1.3157

Martorana, M., Kuhn, T., Stork, L., & Ossenbruggen, J.V. (2024). Zero-Shot Topic Classification of Column Headers: Leveraging LLMs for Metadata Enrichment. International Conference on Semantic Systems. arXiv-Computer Science-Databases. https://doi.org/10.48550/arXiv.2403.00884

Michael, B., & Han, M.J. (2020). Assessing BIBFRAME 2.0: Exploratory Implementation in Metadata Maker. International Conference on Dublin Core and Metadata Applications, 2020. https://doi.org/10.23106/dcmi.952141614

OCLC. (2023). OCLC-supplied macros. Retrieved from https://help.oclc.org/Metadata_Services/Connexion/Connexion_client/Connexion_client_basics/Use_macros/Get_started/50OCLC_supplied_macros

Park, J.R., & Maszaros, S. (2009). Metadata Object Description Schema (MODS) in digital repositories: An exploratory study of metadata use and quality. Knowledge Organization, 36(1), 46–59.https://doi.org/10.5771/0943-7444-2009-1-46

Park, J.R., & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms. Cataloging & Classification Quarterly, 48(8), 696–715. https://doi.org/10.1080/01639374.2010.508711

Riva, P., Lebeouf, P., & Žumer, M. (2017). IFLA Library Reference Model: A Conceptual Model for Bibliographic Information. Retrieved August 12, 2024, from https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. Journal of the American Society of Information Science, 58, 1720–1733. https://doi.org/10.1002/asi.20652

Taniguchi, S. (2024). Creating and evaluating MARC 21 bibliographic records using ChatGPT. Cataloging & Classification Quarterly, 62 (5), 527-546. https://doi.org/10.1080/01639374.2024.2394513

Weagley, J., Gelches, E., & Park, J. (2010). Interoperability and metadata quality in digital video repositories: a study of Dublin Core. Journal of Library Metadata, 10(1), 37–57. https://doi.org/10.1080/19386380903546984

Zavalin, V.I., & Zavalina, O. L. (2023). Exploration of accuracy, completeness, and consistency in metadata for physical objects in museum collections. In, Information for a Better World: Normality, Virtuality, Physicality, Inclusivity: 18th International Conference, iConference 2023, Proceedings (pp. 83-90). https://doi.org/10.1007/978-3-031-28032-0_7

Zavalin, V., Zavalina, O.L., & Miksa, S.D. (2021). Exploration of subject representation and support of Linked Data in recently created library metadata: Examination of most widely held WorldCat bibliographic records. Library Resources and Technical Services, 65 (4), 154-165. https://doi.org/10.5860/lrts.65n4.1544

Zavalin, V., Zavalina, O.L., & Safa, R. (2021). Patterns of subject metadata change in MARC21 bibliographic records representing video recordings. Proceedings of the Association for Information Science and Technology, 58(1). https://doi.org/10.1002/pra2.494

Zavalina, O.L., & Burke, M. (2021). Assessing skill-building in metadata instruction: Quality evaluation of Dublin Core metadata records created by graduate students. Journal of Education for Library and Information Science, 62(4), 423-442. https://doi.org/10.3138/jelis.62-4-2020-0083

Zhang, S., Wu, M., & Zhang, X. (2023). Utilising a large language model to annotate subject metadata: a case study in an Australian national research data catalogue", arXiv-CS-Computation and Language. https://doi.org/10.48550/arXiv.2310.11318

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Scientists, but deny science? Climate change sceptics networks on YouTube led by scientists

Qiaoyi Liu, Yuheun Kim, and Jeff Hemsley

DOI: https://doi.org/10.47989/ir30iConf47212

Abstract

Introduction. Climate change debates have divided our society more than ever. Despite most scientists believing in anthropogenic climate change, a small group of people with scientific knowledge and reasoning are denying it.

Method. In this paper, we collect YouTube video comments' data to study the content posted by climate change sceptical scientists and their impact on comment social networks.

Analysis. We apply natural language processing and social networks analyses to study those comments and networks.

Results. We find that denying scientists question the validity of anthropogenic climate change using objective terms such as '*Co2*', '*history*', '*data*', etc., while non-scientists rarely mention these terms, instead frequently using words like '*money*', '*truth*'. Scientists-led social networks are also more structured with significant core users, while non-scientists-led networks have smaller and fragmented groups, indicating scientists-led discussions on climate change are more stable and consistent.

Conclusions. Scientists who deny human-caused climate change cast greater influence on the climate change denying social networks. Their opinions using more scientific terms cause the networks to be more centralized and form more consistent patterns.

Introduction

Despite the existence of scientifically well-established results, such as that emission of greenhouse gases produces global warming, there are sizable proportions of the population in the US who reject climate science (Lewandowsky & Oberauer, 2016). Disbelief in climate change can mean either the absolute rejection of the existence of anthropogenic climate change (ACC) or a lack of sureness about the anthropogenic cause of climate change i.e. climate change scepticism (Lübke, 2022). There have been efforts to increase the influence of environment interest (Jacques et al., 2008) and advertising campaigns to shape the public discourse about climate change (Kovaka, 2021), yet still, organized climate change deniers persist on social media platforms (Falkenberg et al., 2022; K. M. d'I. Treen et al., 2020; Williams et al., 2015) and are polarized from climate change believers (Björnberg et al., 2017; Dunlap, 2013; Hornsey et al., 2016).

Making decisions about complex scientific topics like climate change requires more than just better knowledge of the facts. It takes the ability to critically evaluate evidence and explanations, take into account the source of that information, and appreciate how the methods of science lead to specific conclusions (Sinatra & Hofer, 2021). Studies show that the disbelief in climate change is because of: (a) lack of *knowledge*, meaning people cannot comprehend scientific consensus or crucial scientific facts; (b) lack of *cognition*, meaning people have poor reasoning; and (c) *social or political identity*, as in admitting climate change may be a betrayal to people's identities (Kovaka, 2021). Accumulating knowledge can also improve one's higher cognitive process (Halford et al., 2010). Therefore, reason (a) and (b) are often interchangeable and simultaneous. For the public, many do not know or misunderstand the scientific process that produces these findings, challenging their ability to evaluate the information posted on social media (Sinatra & Hofer, 2021).

Following this line of thinking, we may assume scientists who have higher education training possess both the knowledge and cognition level to understand and agree to anthropogenic climate change. However, this is not always the case. Studies find that the general level of education, scientific knowledge, and science literacy can only modestly predict one's attitude toward and trust in science (Lewandowsky & Oberauer, 2016). General education and scientific literacy not only mitigate the rejection of science facts such as climate change, but also increase the polarization of opinions (Lewandowsky & Oberauer, 2016). A study on comments by climate change sceptics shows a particularly striking aspects of threads come from people with high level of educational backgrounds (Matthews, 2015).

In this study, we use motivated reasoning theory to study climate change sceptics scientists in the social media networks (Kunda, 1990). We conduct sentiment and content analysis using YouTube data to understand what reasons scientists give to reject human-caused climate change, and their impact on the climate change sceptics' network. We use social network analysis to test whether scientist-led networks differ from non-scientist-led sceptics' network. This paper will first briefly review studies on science denialism and climate change polarization on social media. Next, we propose hypotheses to address our research questions and justify methodologies used in this study. In discussion and conclusion, we interpret the results of analyses and discuss our contributions to the social media climate change research.

Climate change denialism by scientists

Climate change sceptical scientists are generally people who (i) do not accept human-caused climate change, (ii) personally consider themselves to be pro-science, and (iii) are aware that most scientists agree on the reality of human-caused climate change (Kovaka, 2021). Their existence is a puzzle because they are trained with scientific knowledge and have reasoning capabilities enough to understand and appreciate scientific evidence while rejecting ACC. They are well-aware of how science works and what makes it a successful tool for producing knowledge (Kovaka, 2021). Thus, we wonder: what is leading them to disagree with ACC? One reason is that scientists also have

their own political and socio-cultural identities, such as their professional socialization and their hostility against the increasing allocation of government funding to applied research rather than to basic science (Lahsen, 2008). Many of the denying scientists are not affiliated with any academic institution but are paid through working at a think tank (Björnberg et al., 2017). The link between political/socio identity and opinions on climate change has been studied by many researchers. However, the generalizability of this link does not exactly explain how sceptics scientists deny climate change in particular. Reasons (a) and (b) are also not applicable to scientists, since studies show that general science literacy and quantitative skills do not vary between people who accept or reject human-caused climate change (Kahan et al., 2012).

Therefore, we turn to a fourth possible reason for denying human-caused climate change: (d) scepticism of the quality and validity of climate science. Some climate change sceptics are sceptical of climate science as reliable with objective scientific evidence, according to standards of good science (Kovaka, 2021; Matthews, 2015). Driven by motivated reasoning, the misconception of climate science as poor science makes it easy to deny the reality of human-caused climate change, which may cause scientists to use their knowledge and educated privilege to deny ACC. In this study, we are interested in finding out whether that is true. Our aim is to know what exactly sceptics scientists post on social media to deny ACC and what arguments they tend to propose to support this opinion. We expect sceptics scientists to have higher probability using words which indicate scientific research standards, evidence, statistical figures, etc. Therefore:

> *H1: Climate change sceptics scientists deny anthropogenic climate change by questioning the trustworthiness and validity of climate science results.*

Climate change social media network

Social media platforms are gradually replacing the role of traditional media as outlets for communication on crisis such as climate change (Duan et al., 2023). The explosion of social media platforms over the past decade has provided us with rich data, allowing us to study the communication and interactions among people with richer data than ever before (Pearce et al., 2019). Applying traditional social network analysis approach to the study of social media networks, researchers are able to break the conventional boundaries of social networks (Kane et al., 2014) as well as access much larger data. However, because social media platforms are also using algorithms to separate users based on their preference, more commonly we see homogeneous networks with people of common attitudes occupying similar network positions (Erickson, 1988). Climate change polarization on platforms such as Twitter/X, Facebook, and Reddit (Bloomfield & Tillery, 2019; Falkenberg et al., 2022; K. Treen et al., 2022) shows that users on social media are more likely to communicate with others who have shared opinions with them over complex science decisions like climate change.

In climate change sceptics' networks, we intend to study the role of scientists who deny or are sceptical of climate change and their impact on the structure of the network around their content. Scientists who are climate change sceptics receive a large amount of media attention and wield significant influence in the societal debate about climate change impacts and policy (Anderegg et al., 2010). However, little is known about this impact to social networks on social media platforms. Longitudinal studies on the climate change denier networks exhibits high stability in terms of its core leaders. The overall size of the network remains relatively stable over time, while having small groups within the climate change deniers' network fluctuate over time. A high number of small groups in a network can lead to fragmentation and reduced interactions across the whole network (Yang, 2024). Therefore, we are interested in how the involvement of scientists may or may not influence the structure of the climate change sceptics' networks on social media compared to non-scientists-led networks. We propose:

H2: Climate change sceptics networks led by scientists present different patterns compared to networks led by non-scientists.

Methods

We use YouTube as our data source to test our hypotheses. Studies show that YouTube's recommender algorithm heightens individual political opinions' polarization (Cho et al., 2020) and users on YouTube especially prioritize conspiracy theories over scientific mainstream views (Allgaier, 2019; Bessi et al., 2016). The combination of video and text contents together can also lead to higher levels of engagement than text alone (Yadav et al., 2011). Video posts create higher structural virality than images and news stories (Juul & Ugander, 2021). Therefore, we use YouTube data to test our hypotheses.

Data

We selected the ten most-viewed YouTube videos using the search term 'climate hoax scientist' for scientist-led videos and fifteen videos using the search term 'climate hoax' for non-scientist-led videos. Comments are collected using the YouTube API and *vosonSML* package in R (Gertzel et al., 2022). The '*climate hoax scientist*' videos are created by scientists or scientists' interviews or speeches with a doctoral degree according to their academic personal webpages. The '*climate hoax*' search term represents the general climate change sceptics videos, usually created by influencers. The total number of comments from scientists' videos is 64,604 and the total number of comments from non-scientists' videos is 230,103. We excluded irrelevant videos that are about specific organizations/conferences/activists, education materials, children's lyrics, etc. We use the two types of comments data separately to test our hypotheses on scientists' claims and social networks.

We also collected the transcripts of all 25 videos to conduct content analysis. We are aware that this data only represents a partial scope of the climate change denial discussion space. Our goal is to explore the phenomena of climate change scepticism by scientists and discuss their impact on climate change sceptics' network.

Natural language processing (NLP) and sentiment analysis

We identified significant keywords from the video transcripts and calculated sentiments of the comments data using natural language processing (NLP) techniques. First, we cleaned the comments data by removing URLs and user mentions. In identifying significant keywords in the comments, we extracted nouns from the transcript data using the NLTK package (Bird et al., 2009) and drew it as a word cloud.

For sentiment analysis, we fine-tuned DistilBERT (Sanh et al., 2019), a distilled version of the BERT model, using Twitter data related to climate change sentiments (Twitter). This ensures the model would incorporate relevant knowledge for sentiment analysis. The training data consisted of tweets annotated with four labels: news, positive, neutral and negative. For our analysis, we excluded the news label focusing on the remaining three sentiment labels. The dataset was split in a ratio of 80:20 for training and testing.

We trained DistilBERT with a batch size of 16 over two epochs, using a learning rate of 2e-5. The model achieved an F1-score of approximately 0.78, indicating decent performance. Using the fine-tuned model, we inferenced our YouTube comments dataset to calculate the probability score for each sentiment.

Social network analysis (SNA)

To explore the network structures of climate change sceptics videos, we conducted SNA on the user comment data for the two types of videos we collected (referred as scientists network and non-scientists network). We created edgelists to represent the relations between users. One key

issue in this process is expanding the levels of nested comments. YouTube limits the number of nested comments from exceeding two layers, shown in figure 1. Expanding the second layer of nested comments gives us higher degrees of the networks. Thus, our first step is to extract usernames from the nested comments and assign them to the lower level, which would indicate one user referring to a previous user's comment (see figure 1). Next, we use the *igraph* package in R to generate the network plots (Csardi & Nepusz, 2006). For the plots, we used the stress-minimization algorithm known as the Kamada-Kawai (KK) layout (Kamada & Kawai, 1989). In this layout, the spatial distance between nodes is longer (shorter) when the shortest path those two nodes is longer (shorter). Note that we have excluded the nodes for videoIDs and their edges as well as any self-commenting edges for improved visualization.

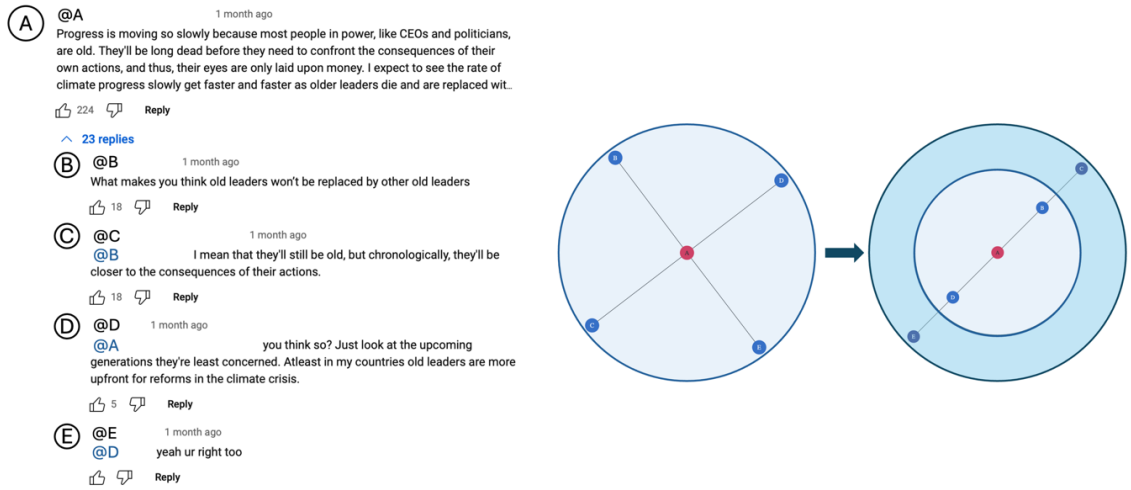


Figure 1. Example of how we expanded the nested comments to retrieve the original degrees of the commenting networks

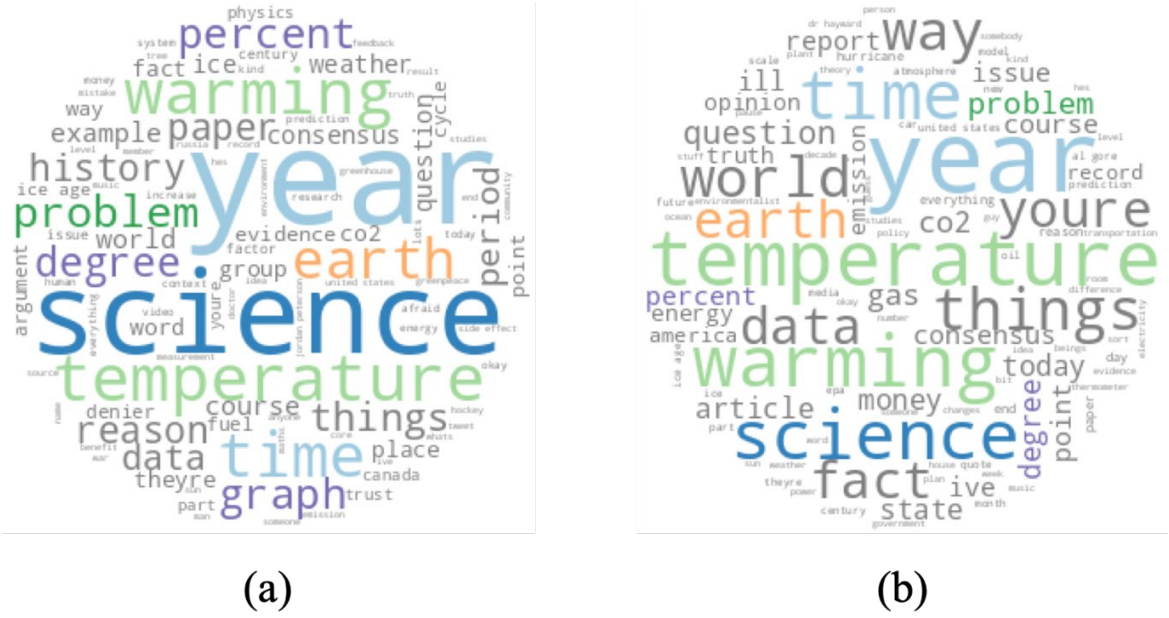


Figure 2. Word cloud of the most mentioned nouns from YouTube video transcript. Word cloud (a) is from skeptic scientists videos and word cloud (b) is from non-scientists videos.
Word size reflects frequency of mention.

Results

The word cloud of '*climate hoax scientist*' videos (figure 2a) emphasize the words '*science*' and other terms that convey objectivity, such as '*graph*', '*percent*', and '*degree*', more frequently than '*climate hoax*' videos (figure 2b). This suggests that these videos focus more on scientific facts. Words such as '*co2*', '*history*', and '*ice age*' further support the stance of scientists denying anthropogenic climate change, at the same time using data-driven language to back their opinions. The use of these terms supports H1, as scientists question the validity of climate science results.

On the other hand, '*climate hoax*' videos frequently mention intuitive weather-related terms, such as '*temperature*' and '*warming*'. Notably, while '*truth*' is mentioned often, objective terms are used less frequently compared to the '*climate hoax scientist*' videos. Another unique aspect is the use of the word '*money*', possibly highlighting economic factors associated with climate change.

We used the results of sentiment analysis on the social network plots – blues nodes are positive, red nodes are negative, gray nodes are neutral, and green nodes are sentiment unknown. Both scientists and non-scientists network show highly homogeneous positive sentiments (see table 1), indicating most users agree with climate change denialism. Scientists-led networks show a significant core at the centre with dense edges between the nodes in the centre. There are also nodes outside the core but form circle layers. This is due to the KK layout which optimizes the space of the plot such that the nodes are spaced farther apart if the path in the network between two nodes are longer. Two nodes are closer together if there are fewer hops between nodes connecting them, so nodes that are on the same '*circle*' are ones that share equal number of '*hops*' to reach other nodes. Whereas the non-scientists-led networks are messier, with less significant core and having more outliers isolated from the rest of the network. This may imply these non-science users are less connected to other users, which means they could be complaining about climate change rather than having meaningful discussions over the issue. The scientists-led networks are more structured, which confirms our H2 that scientists are commenting on climate change negatively but with an understanding of science.

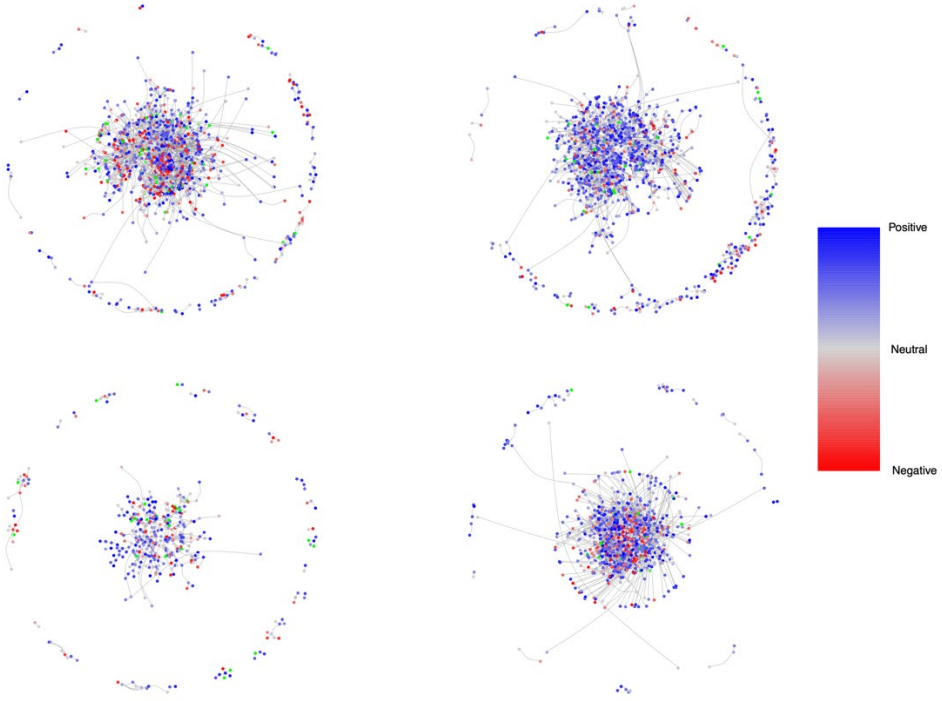


Figure 3. Scientists-led network (green nodes are sentiment unknown)

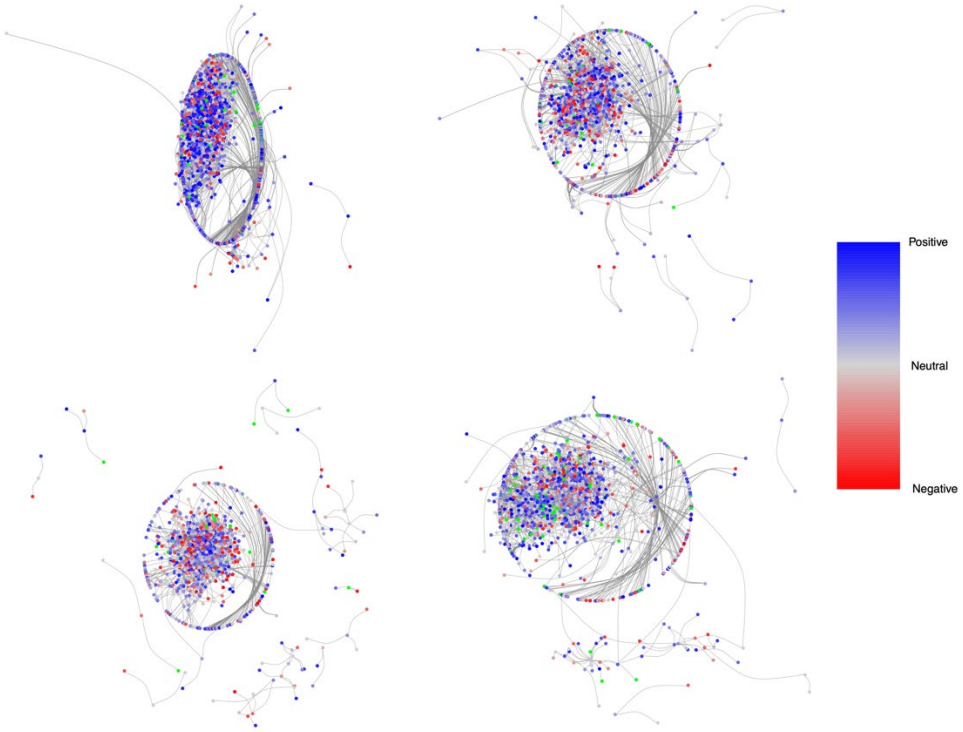


Figure 4. Non-scientists-led network (green nodes are sentiment unknown)

Network Type	Positive Sentiment Overall Probability	Negative Sentiment Overall Probability	Neutral Sentiment Overall Probability
Scientists network	53.03%	21.18%	25.79%
Non-scientists network	51.89%	22.62%	25.49%

Table 1. Sentiment of scientists and non-scientists networks

Discussion

The results of our study confirm our hypotheses that scientists deny climate change by questioning whether climate science meets the standards of good science inquiries (*H1*), and they lead social networks perform different structures from non-scientist led networks (*H2*). Their objection to some climate change evidence as being invalid is quite popular among deniers. This could be due to the general social media user not having sufficient understanding of how science is really conducted and how it differs from what they imagined it to be (Kovaka, 2021). Influenced by motivated reasoning, they are more likely to believe in claims saying climate scientists and their results are not objective but biased and based on self-interests. The spread of misinformation and disinformation about science, magnified by social media algorithms and echo chambers is creating more scepticism and mistrust (Sinatra & Hofer, 2021).

The conflict over how to address climate change can be seen as a symbolic political and cultural struggle over the dominant cultural understanding of this issue (Brulle, 2021). The rejection of this environmental issue derived from political, cognitive, and identity concerns can be seen as an objection and rebellion against science as the privileged epistemic position in society (Kovaka, 2021). With scientists who belong to this privileged community taking the lead in the organized

scepticism of anthropogenic climate change, we see a tighter social network than networks led by non-scientists such as influencers. This means that scientists still play an important role in influencing public opinions over complex scientific issues. The scepticism on anthropogenic climate change expressed by scientists probably increases polarization around climate change. However, we do not believe having such diverse and disagreeing opinions is entirely unbeneficial to society. '*There is nothing more anti-scientific than the very idea that science is settled, static, impervious to challenge.*' (Krauthammer, 2014). This idea urges climate science research communities to provide more solid and transparent evidence of their scientific process to convince the public who are temporarily sceptical only due to the lack of trustworthy results.

Conclusion

We study the involvement of climate change sceptical scientists on social media and their impact on social networks of commenting users. We find that these scientists are inclined to use more scientific terms to deny anthropogenic climate change, casting a greater influence on the social network structure. Compared to non-scientists' networks, these networks are more centralized and form more consistent patterns. One limitation of this study is that we only focus on a segment of the large social media data available on a particular group of subjects. Further research is required to fully understand the activities and impact of scientists on anthropogenetic climate change scepticism and societal science denialism.

About the authors

Qiaoyi Liu is a Ph.D. student in Information Science and Technology at Syracuse University. Her research interests are in knowledge organization (KO) and science of science (SoS). Especially, she studies biological knowledge representation and construction of ontologies guided by classification theories and semantic measurements. She is interested in knowledgebases exploited by ML models and trustworthy LLMs to generate knowledge for bioinformatics and computational biology research.

Yuheun Kim is a Ph.D. student in Information Science and Technology at Syracuse University. Her research interest is computationally approaching stereotypes and bias across different cultures using natural language processing (NLP) tools.

Jeff Hemsley is the Interim Dean and an Associate Professor at the School of Information Studies at Syracuse University. He earned his Ph.D. from the University of Washington's information school, and his research is about understanding information diffusion on social networks. He is particularly interested in viral information events and is co-author of the book Going Viral (Polity Press, 2013 and winner of ASIS&T best science books of 2014 Information award and selected by Choice magazine as an outstanding academic title for 2014), which explains what virality is, how it works technologically and socially, and draws out the implications of this process for social change.

References

Allgaier, J. (2019). Science and Environmental Communication on YouTube: Strategically Distorted Communications in Online Videos on Climate Change and Climate Engineering. Frontiers in Communication, 4. https://doi.org/10.3389/fcomm.2019.00036

Anderegg, W. R. L., Prall, J. W., Harold, J., & Schneider, S. H. (2010). Expert credibility in climate change. Proceedings of the National Academy of Sciences, 107(27), 12107–12109. https://doi.org/10.1073/pnas.1003187107

Bessi, A., Zollo, F., Vicario, M. D., Puliga, M., Scala, A., Caldarelli, G., Uzzi, B., & Quattrociocchi, W. (2016). Users Polarization on Facebook and Youtube. PLOS ONE, 11(8), e0159641. https://doi.org/10.1371/journal.pone.0159641

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc.

Björnberg, K. E., Karlsson, M., Gilek, M., & Hansson, S. O. (2017). Climate and environmental science denial: A review of the scientific literature published in 1990–2015. Journal of Cleaner Production, 167, 229–241. https://doi.org/10.1016/j.jclepro.2017.08.066

Bloomfield, E. F., & Tillery, D. (2019). The Circulation of Climate Change Denial Online: Rhetorical and Networking Strategies on Facebook. Environmental Communication, 13(1), 23–34. https://doi.org/10.1080/17524032.2018.1527378

Brulle, R. J. (2021). Networks of Opposition: A Structural Analysis of U.S. Climate Change Countermovement Coalitions 1989–2015. Sociological Inquiry, 91(3), 603–624. https://doi.org/10.1111/soin.12333

Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization. Journal of Broadcasting & Electronic Media, 64(2), 150–172. https://doi.org/10.1080/08838151.2020.1757365

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. In InterJournal: Vol. Complex Systems (p. 1695). https://igraph.org

Duan, Y., Khoury, C., Smith, A. O., Joh, U., & Hemsley, J. (2023). Comparing Crisis Communication on TikTok and YouTube: A Case Study of the 2023 California Floods. Proceedings of the Association for Information Science and Technology, 60(1), 949–951. https://doi.org/10.1002/pra2.908

Dunlap, R. E. (2013). Climate Change Skepticism and Denial: An Introduction. American Behavioral Scientist, 57(6), 691–698. https://doi.org/10.1177/0002764213477097

Erickson, B. H. (1988). The relational basis of attitudes. In Social structures: A network approach. (pp. 99–121). Cambridge University Press.

Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrociocchi, W., & Baronchelli, A. (2022). Growing polarization around climate change on social media. Nature Climate Change, 12(12), 1114–1121. https://doi.org/10.1038/s41558-022-01527-x

Gertzel, B., Borquez, F., Graham, T., & Ackland, R. (2022). vosonSML: Collecting Social Media Data and Generating Networks for Analysis. VOSON Lab, Virtual Observatory for the Study of Online Networks. https://github.com/vosonlab/vosonSML

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. Trends in Cognitive Sciences, 14(11), 497–505. https://doi.org/10.1016/j.tics.2010.08.005

Hornsey, M. J., Harris, E. A., Bain, P. G., & Fielding, K. S. (2016). Meta-analyses of the determinants and outcomes of belief in climate change. Nature Climate Change, 6(6), 622–626. https://doi.org/10.1038/nclimate2943

Jacques, P. J., Dunlap, R. E., & Freeman, M. (2008). The organisation of denial: Conservative think tanks and environmental scepticism. Environmental Politics. https://doi.org/10.1080/09644010802055576

Juul, J. L., & Ugander, J. (2021). Comparing information diffusion mechanisms by matching on cascade size. Proceedings of the National Academy of Sciences, 118(46), e2100786118. https://doi.org/10.1073/pnas.2100786118

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. Nature Climate Change, 2(10), 732–735. https://doi.org/10.1038/nclimate1547

Kamada, T., & Kawai, S. (1989). AN ALGORITHM FOR DRAWING GENERAL UNDIRECTED GRAPHS. INFORMATION PROCESSING LETTERS, 31(1).

Kane, G. C., Alavi, M., Labianca, G. (Joe), & Borgatti, S. P. (2014). What'S Different About Social Media Networks? A Framework and Research Agenda. MIS Quarterly, 38(1), 275–304.

Kovaka, K. (2021). Climate change denial and beliefs about science. Synthese, 198(3), 2355–2374. https://doi.org/10.1007/s11229-019-02210-z

Krauthammer, C. (2014, February 21). Opinion | The myth of 'settled science.' Washington Post. https://www.washingtonpost.com/opinions/charles-krauthammer-the-myth-of-settled-science/2014/02/20/c1f8d994-9a75-11e3-b931-0204122c514b_story.html

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480.

Lahsen, M. (2008). Experiences of modernity in the greenhouse: A cultural analysis of a physicist "trio" supporting the backlash against global warming. Global Environmental Change, 18(1), 204–219. https://doi.org/10.1016/j.gloenvcha.2007.10.001

Lewandowsky, S., & Oberauer, K. (2016). Motivated Rejection of Science. Current Directions in Psychological Science, 25(4), 217–222. https://doi.org/10.1177/0963721416654436

Lübke, C. (2022). Socioeconomic Roots of Climate Change Denial and Uncertainty among the European Population. European Sociological Review, 38(1), 153–168. https://doi.org/10.1093/esr/jcab035

Matthews, P. (2015). Why Are People Skeptical about Climate Change? Some Insights from Blog Comments. Environmental Communication, 9(2), 153–168. https://doi.org/10.1080/17524032.2014.999694

Pearce, W., Niederer, S., Özkula, S. M., & Sánchez Querubín, N. (2019). The social media life of climate change: Platforms, publics, and future imaginaries. WIREs Climate Change, 10(2), e569. https://doi.org/10.1002/wcc.569

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019, October 2). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv.Org. https://arxiv.org/abs/1910.01108v4

Sinatra, G., & Hofer, B. (2021). Science Denial: Why It Happens and What to Do About It (1st ed.). Oxford University PressNew York. https://doi.org/10.1093/oso/9780190944681.001.0001

Treen, K. M. d'I., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. WIREs Climate Change, 11(5), e665. https://doi.org/10.1002/wcc.665

Treen, K., Williams, H., O'Neill, S., & Coan, T. G. (2022). Discussion of Climate Change on Reddit: Polarized Discourse or Deliberative Debate? Environmental Communication, 16(5), 680–698. https://doi.org/10.1080/17524032.2022.2050776

Twitter Climate Change Sentiment Dataset, https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset/data, last accessed 2024/09/09

Williams, H. T. P., McMurray, J. R., Kurz, T., & Hugo Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. Global Environmental Change, 32, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Yadav, A., Phillips, M. M., Lundeberg, M. A., Koehler, M. J., Hilden, K., & Dirkin, K. H. (2011). If a picture is worth a thousand words is video worth a million? Differences in affective and cognitive processing of video and text cases. Journal of Computing in Higher Education, 23(1), 15–37. https://doi.org/10.1007/s12528-011-9042-y

Yang, A. (2024). *Exploring the Network and Topic Stability in Climate Change Deniers' Disinformation Network: A Longitudinal Study*. https://hdl.handle.net/10125/106714

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Uncovering strategies for identifying deepfakes

Celene Neo, Dion Hoe-Lian Goh, Chun Wan Ying Rachel, and Chei Sian Lee

DOI: https://doi.org/10.47989/ir30iConf47209

Abstract

Introduction. The proliferation of generative artificial intelligence tools capable of producing high-quality videos that can masquerade as genuine content has raised concerns about online misinformation. This study investigates human ability to identify deepfake videos, with a focus on identification performance and the strategies employed.

Method. Data was collected through an online survey. Participants were young adults aged 21 to 35. They were shown four videos and asked to identify them as real or deepfake, followed by questions about the identification strategies used.

Results. Our results revealed the diverse range of strategies utilised. Predominant strategies centre around assessing the authenticity of traits pertaining to the video's subject as opposed to peripheral details. Furthermore, we uncovered preferences for intuition and strategies that relate to individual decision-making over consulting other individuals or online materials.

Conclusion. Our results help enhance understanding of how people identify deepfake videos, adding to existing knowledge. These findings also inform initiatives aimed at educating the public about spotting deepfakes.

Introduction

While generative artificial intelligence (AI) has opened a wealth of creative possibilities, its ability to produce convincing misinformation has made it harder to discern artificial content from real content (Park, 2024). In recent times, misinformation has taken the form of deepfakes – artificially generated videos that manipulate individuals to appear to say or do things they never did (Somers, 2020). Deepfakes have gained notoriety in the public sphere primarily for their use in generating pornographic content, and more recently, to erode trust by creating high-quality, fabricated videos that falsely depict influential figures making controversial statements.

Existing studies (Heidari et al., 2024; Jung et al., 2020; Pan et al., 2020) on deepfake detection are primarily focused on those involving deep learning. Few studies investigate human deepfake detection strategies. Among those that do, they are often limited in terms of sample size and generalisability due to the use of qualitative data collection methods (Goh et al., 2022; Zeng et al., 2023). Videos used in these studies also lack diversity and consist largely of entertainment or political videos (Goh, 2024), which may not sufficiently encompass the genres of videos that participants encounter day-to-day.

This study addresses these gaps by determining human ability to identify deepfake videos through an online quantitative survey. The first objective ascertains deepfake and real video identification performance; and the second examines the strategies that people employ to identify deepfake and real videos.

Literature review

Deepfakes pose significant threats to society, particularly through their potential to create fabricated videos for spreading misinformation. The misuse of deepfakes risks undermining public trust and has the potential to deepen social divisions (Westerlund, 2019). The impact of deepfakes is further amplified when considering that individuals generally perceive video content as more credible than textual information (Sundar et al., 2021), making the spread of falsehoods through deepfakes particularly damaging.

Given the rise in deepfakes (Sumsub, 2023), there is a wealth of research on deepfake detection using deep learning models (El-Gayar et al., 2024; Lee et al., 2023). These approaches enable the rapid processing of large volumes of video data while providing objective assessments (Heidari et al., 2024), which may explain the heightened research interest over human detection capabilities.

While there is a growing body of work on human deepfake detection strategies, data collection methods are often limited to exploratory and qualitative modes such as interviews (Goh, 2024) or diary studies (Zeng et al., 2023). Despite yielding detailed insights, they are constrained by small sample sizes and limited generalisability. Findings are also often limited to the most frequently used strategies regardless of video type. For instance, in Goh (2024), strategies for identifying video authenticity are described without differentiation between methods used for detecting deepfakes or real videos. This approach limits the ability to understand which specific strategies are most effective for each type of video.

Methodology

Our study utilized an online survey. Participants were shown four videos — two real and two deepfake — randomly selected from a pool of ten authentic and ten deepfake videos. These videos were publicly available from the Web and covered a range of topics, including entertainment, politics, education, and sports. Video descriptions can be found in Table 1.

After viewing each video, participants were asked to: (1) identify whether the video was real or deepfake, (2) report their confidence level, and (3) select the strategies they used to arrive at their decision. The strategies were categorized into three types: visual (e.g., facial features, background,

and environmental issues), auditory (e.g., vocal features, sound quality), and knowledge-based (e.g., online tools, knowledge of video subject). These strategies were adapted from prior research (Goh et al., 2022; Zeng et al., 2023; Goh, 2024). Participants could select multiple strategies.

Young adults aged 21 to 35 were recruited as they represent a group that is active online and would likely have experience with deepfake videos (Petrosyan, 2024). A total of 195 participants were recruited via convenience and snowball sampling.

Video Type	Topic	Description
Deepfake	Entertainment	Mark Zuckerberg says he controls billions of peoples' confidential data and thus owns their future.
		Kim Kardashian tells how she likes making money by manipulating her fans.
		Tom Cruise's daily life.
		The Shining movie clip.
	Politics	Manoj Tiwari criticized an opposing political party and encouraged people to vote for his party.
		Jeremy Corbyn supports Boris Johnson as Prime Minister.
		A speech for the Apollo 11 mission gone wrong.
	Educational	Obama reminds people to be more alert to fake news.
		Morgan Freeman asks people: Is seeing believing?
	Sports	Jose Mourinho comments on soccer.
Real	Entertainment	Mark Zuckerberg says he could ascertain people's online behaviours.
		Kim Kardashian claims that she cheated on an exam.
		Tom Holland taking a break from social media.
		Another *The Shining* movie clip.
	Politics	Biden criticized MAGA Republicans.
		Trump blamed congressional attackers and told his supporters to calm down.
		President Uhuru mourns former Kenyan President Mwai Kibaki.
	Educational	Hillary Clinton talks about fake news dangers.
		Ellen warns people about fake news.
	Sports	Jose Mourinho on Sir Alex Ferguson's response after Porto's Champions League win over Manchester United.

Table 1. Description of videos used in the study

Results

Participant demographics

Participants comprised 83 males and 112 females aged 21 to 35 years. The majority were from fields such as social sciences, finance, engineering, sciences, and computing. YouTube was the most used video platform among participants, followed by Instagram and TikTok. A majority (86%) watched videos daily, with most coming across dubious content '*once in a while*' when watching videos.

Video identification performance

To address the first objective, we found that the distribution of participants skewed towards a higher number of correct identifications. Of 195 participants, the majority (73%) were able to identify more than half of the videos correctly. Only 1% identified zero videos correctly, 6%

correctly identified one video, 21% correctly identified two videos, 34% correctly identified three videos, and 38% correctly identified all four videos (Figure 1).

Figure 2 shows identification accuracy by authenticity type. Here, 53% of participants were able to correctly identify both deepfake videos, while 65% correctly identified both real videos.

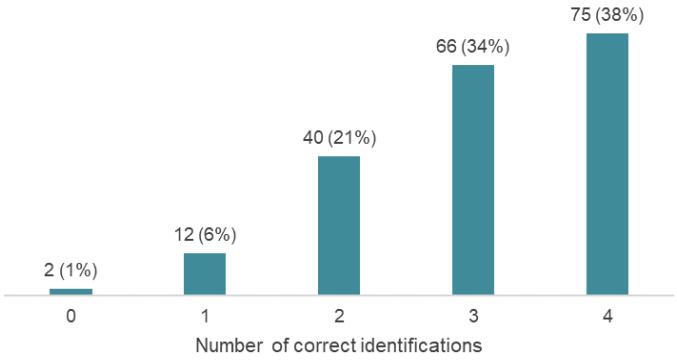


Figure 1. Frequency of number of videos identified correctly

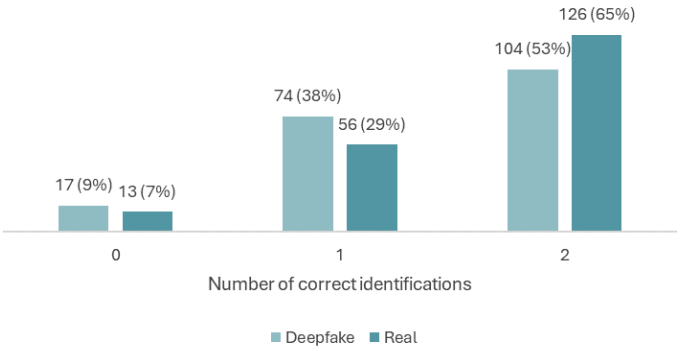


Figure 2. Frequency of correct identifications by authenticity type

Strategies used in deepfake video identification

Table 2 shows the five most and least used strategies associated with correct identification of deepfake videos. The most frequently used methods typically focused on examining the characteristics of the subject in the video, while the least used methods focused on the peripheral details of the video. Among the most used methods, physical and behavioural characteristics, and intuition and emotions were used half of the time.

Note that the frequency values presented in the tables reflect the total number of times each strategy was employed across both deepfake videos watched (Tables 2 and 3) as well as real videos (Tables 4 and 5). Each of the 195 participants viewed two real and two deepfake videos, resulting in 390 video assessments per video type. The percentages for each strategy are calculated using this total number of video assessments as the base.

Table 3 shows the top and bottom five strategies associated with incorrect identification of deepfake videos. Strategies were similar to those used in the correct identification of deepfake videos, except for intelligibility and language, and knowledge of video content which were not present in Table 2.

Strategy		Frequency
Most used	Physical and behavioural characteristics (*Analysing individuals' expressions, body gestures, and movements in the video.*)	203 (52.1%)
	Intuition and emotions (*Based on own instinct, opinions, or emotions.*)	202 (51.8%)
	Vocal features (*Assessing modulation and naturalness in the speaker's voice.*)	185 (47.4%)
	Facial features (*Analysing facial characteristics, such as skin tone, facial symmetry, and hairstyle.*)	173 (44.4%)
	Knowledge of person (*Prior knowledge of people in video.*)	124 (31.8%)
Least used	Colour and lighting inconsistencies (*Evaluating the video's lighting conditions.*)	76 (19.5%)
	Production issues (*Editing issues, camera angle/work issues, shakiness, and jitter.*)	67 (17.2%)
	Background sound issues (*Background reverberation/echoes, overall noise, mechanical noises.*)	66 (16.9%)
	Use of multiple sources (*Consulting multiple sources when using online tools or communicating with others.*)	57 (14.6%)
	Communication with others (*Checking with family or friends either offline or online.*)	53 (13.6%)

Table 2. Strategies associated with correct deepfake identification

Strategy		Frequency
Most used	Vocal features	86 (22.1%)
	Intuition and emotions	72 (18.5%)
	Facial features	69 (17.7%)
	Physical and behavioural characteristics	68 (17.4%)
	Intelligibility and language (*Assessing the language used, fluency, pronunciation, and intelligibility of speech.*)	66 (16.9%)
Least used	Knowledge of video content (*Familiarity with events in the video.*)	32 (8.2%)
	Production issues	31 (7.9%)
	Colour and lighting inconsistencies	30 (7.7%)
	Use of multiple sources	25 (6.4%)
	Communication with others	21 (5.4%)

Table 3. Strategies associated with incorrect deepfake identification

Strategies used in real video identification

Table 4 shows the five most and least used strategies associated with correct identification of real videos. The strategies aligned with those used to correctly identify deepfakes, with the only differing strategy being intelligibility and language.

In terms of the most used strategies, it was interesting to observe that vocal features and facial features saw higher frequency of use as compared to strategies associated with correct identification of deepfake videos. For example, use of vocal features was the predominant strategy, with a frequency of nearly 70%. Compared to Table 2, intuition and emotions fell from the second

to fifth most used strategy despite maintaining a similar frequency of use. A similar change was observed for physical and behavioural characteristics.

Strategy		Frequency
Most used	Vocal features	267 (68.5%)
	Facial features	206 (52.8%)
	Physical and behavioural characteristics	199 (51.0%)
	Intelligibility and language	190 (48.7%)
	Intuition and emotions	190 (48.7%)
Least used	Online tools (*Using search engines like Google or social media platforms like Facebook and YouTube.*)	94 (24.1%)
	Production issues	76 (19.5%)
	Colour and lighting inconsistencies	66 (16.9%)
	Use of multiple sources	52 (13.3%)
	Communication with others	42 (10.8%)

Table 4. Strategies associated with correct real video identification

Table 5 shows the five most and least used strategies associated with incorrect identification of real videos. Once again, strategies were similar to those used in the correct identification of real videos, which is a common observation between correct and incorrect identification of deepfakes. Notably, this was the only instance where background and environmental details were among the most used strategies.

Strategy		Frequency
Most used	Physical and behavioural characteristics	49 (12.6%)
	Vocal features	47 (12.1%)
	Intuition and emotions	41 (10.5%)
	Background and environmental details (*Scene settings, unusual backgrounds, issues with watermarks, logos, or subtitles.*)	33 (8.5%)
	Facial features	31 (7.9%)
Least used	Online tools	20 (5.1%)
	Knowledge of video content	19 (4.9%)
	Colour and lighting inconsistencies	17 (4.4%)
	Communication with others	17 (4.4%)
	Use of multiple sources	12 (3.1%)

Table 5. Strategies associated with incorrect real video identification

Discussion

Overall, participants demonstrated an ability to differentiate between deepfake and authentic videos. When examining performance based on authenticity type, participants exhibited stronger proficiency in detecting real videos over deepfakes.

Across all videos watched, participants used a variety of strategies. This highlights the difficulty in human detection as there is no singular strategy that individuals can rely on to accurately distinguish deepfakes from real videos (Goh, 2024; Groh, 2020). This finding thus highlights the importance of media literacy and the use of multiple methods to ascertain the authenticity of videos.

Participants largely focused on the subjects in the video, as seen by facial features, and physical and behavioural characteristics emerging as frequently used strategies. In contrast, strategies that considered general video and audio attributes such as production quality were less frequently

utilised. This may stem from the understanding that the subject is often the main focus of deepfake manipulation (Sundar et al., 2021), prompting participants to ignore other details (Groh et al., 2021). The danger however is that as the quality of deepfake manipulation increases, people relying primarily on such strategies may fail to spot falsified content.

The strategies used for identifying deepfake videos differed greatly from those used for image and text identification. A study on deepfake images found that participants often relied on peripheral features, such as accessories or clothing texture, to identify manipulated images (Bray et al., 2023). Similarly, strategies for evaluating the credibility of blog posts emphasised the importance of arrangement and alignment features (Jo et al., 2019). In contrast, deepfake video detection requires the analysis of dynamic components, such as facial expressions and body movement, which adds a layer of complexity compared to static media types. This distinction illustrates the need for a more nuanced approach to video detection, focusing on dynamic strategies that are not as relevant in text or image analysis.

Vocal features were the most used strategy in the identification of real videos. This suggests that participants perceive the human voice as a convincing feature associated with real videos. The complexity of the human voice and the difficulty of deepfakes to replicate these complexities makes vocal features a critical differentiator (Kulangareth et al., 2024), aiding participants in distinguishing between real and fabricated content. However as with the concern about visual features, deepfake technology is increasingly able to accurately clone voices (Mai et al., 2023). Reliance on vocal features only would again pose misidentification dangers.

Notably, regardless of video type and the accuracy of identification, the two least used strategies were employing multiple sources for verification and communication with others. Preference for individual decision-making may be explained by confidence in individual ability to identify deepfakes (Köbis et al, 2021). However, as pointed out by Goh (2024), the use of more cognitively demanding strategies such as referencing multiple sources increases the likelihood of correct identifications. This is especially important due to the rapid advances in deepfake generation technology.

Conclusion

This study reveals the strategies used in deepfake video identification, addressing a gap in the current literature. The findings underscore the complexity of detecting deepfakes, as evidenced by the array of identification methods utilized.

Our study offers theoretical contributions to the field of human deepfake detection and more generally, information credibility assessment, by revealing the strategies young adults use. In contrast to existing research which often reported on most used strategies (Goh et al., 2022; Zeng et al., 2023; Goh, 2024), our larger participant sample allowed for a detailed analysis of both most and least used ones. The exploration of lesser-preferred strategies offers a more comprehensive understanding of identification approaches and their potential shortcomings. Our findings provide a nuanced, multi-dimensional perspective that enhances existing literature on deepfake detection.

In terms of practical implications, our findings can inform educators in the development of media literacy curricula aimed at enhancing digital wellness and safety. Additionally, authorities and online platforms can use these insights to craft targeted strategies for combating deepfakes, including public service announcements, improving resilience against misinformation.

Despite the insights uncovered, several limitations should be acknowledged. First, the strategies recorded in the survey may not reflect the full suite of those employed by individuals in real world scenarios. Participants were informed that they had the option to rewatch videos multiple times, potentially leading to heightened scrutiny than typical online encounters, inflating the accuracy rates. Future research should aim to capture more naturalistic viewing behaviours. Second, as

deepfakes become increasingly sophisticated, new detection strategies may emerge while existing ones may become less effective. Ongoing research is thus essential in providing insights into these evolving methods, ensuring that detection strategies remain relevant and effective in countering ever advancing deepfake technologies.

Acknowledgements

This research project was supported by Nanyang Technological University under the URECA Undergraduate Research Programme as well as a Ministry of Education (Singapore) Tier 2 grant (T2EP40122-0015).

About the authors

Celene Neo is a final-year Communication Studies student at Nanyang Technological University's Wee Kim Wee School of Communication and Information in Singapore. She can be contacted at cneo018@e.ntu.edu.sg.

Dion H. Goh is a Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. His research interests include social media practices and perceptions, game-based techniques for shaping user perceptions and motivating behaviour, and online information sharing and seeking. He can be contacted at ashlgoh@ntu.edu.sg.

Rachel Chun is a Research Associate at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. She is dedicated to extending care and respect to the community through her research and professional endeavours. She can be contacted at rachel.chunwy@ntu.edu.sg.

Chei Sian Lee is a Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University in Singapore. Her research focuses on generative AI and digital nudging, enhancing engagement, decision-making, and learning with an ethical, information-oriented approach. She can be contacted at leecs@ntu.edu.sg.

References

Bray, S. D., Johnson, S. D., Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. Journal of Cybersecurity, 9(1). https://doi.org/10.1093/cybsec/tyad011

El-Gayar, M.M., Abouhawwash, M., Askar, S.S., & Sweidan, S. (2024). A novel approach for detecting deep fake videos using graph neural network. J Big Data 11, 22. https://doi.org/10.1186/s40537-024-00884-y

Goh, D.H., Lee, C.S., Chen, Z., Kuah, X.W., & Pan, Y.L. (2022). Understanding users' deepfake video verification strategies. Communications in Computer and Information Science, 1655. https://doi.org/10.1007/978-3-031-19682-9_4

Goh, D. (2024). He looks very real: Media, knowledge, and search-based strategies for deepfake identification. Journal of the Association for Information Science and Technology; 75(6):643-654. https://doi.org/10.1002/asi.24867

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021). Deepfake detection by human crowds, machines, and machine-informed crowds. Psychological and Cognitive Sciences, 119(1). https://doi.org/10.1073/pnas.2110013119

Groh, M. (2020). Detect DeepFakes: How to counteract misinformation created by AI. MIT Media Lab. https://www.media.mit.edu/projects/detect-fakes/overview/

Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. WIREs Data Mining and Knowledge Discovery, 14(2), e1520. https://doi.org/10.1002/widm.1520

Jo, Y., Kim, M., Han, K. (2019). How Do Humans Assess the Credibility on Web Blogs: Qualifying and Verifying Human Factors with Machine Learning. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). https://doi.org/10.1145/3290605.3300904

Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. IEEE Access, 8, 83144-83154. https://doi.org/10.1109/ACCESS.2020.2988660

Köbis, N. C., Dolezalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. iScience, 24(11), Article 103364. https://doi.org/10.1016/j.isci.2021.103364

Kulangareth, N.V., Kaufman, J., Oreskovic, J., & Fossat, Y. (2024). Investigation of Deepfake Voice Detection Using Speech Pause Patterns: Algorithm Development and Validation. JMIR Biomed Eng, 9. https://doi.org/10.2196/56245

Lee, E. G., Lee, I., & Yoo, S. B. (2023). ClueCatcher: Catching Domain-Wise Independent Clues for Deepfake Detection. Mathematics. 2023; 11(18):3952. https://doi.org/10.3390/math11183952

Mai, K. T., Bray, S., Davies, T., Griffin L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. PLoS One, 18(8), e0285333. https://doi.org/10.1371/journal.pone.0285333

Pan, D., Sun, L., Wang, R., Zhang, X., & Sinnott, R.O. (2020). Deepfake Detection through Deep Learning. 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT),134 - 143. https://doi.org/10.1109/BDCAT50828.2020.00001

Park, H. J. (2024). The rise of generative artificial intelligence and the threat of fake news and disinformation online: Perspectives from sexual medicine. Investig Clin Urol, 65(3), 199-201. https://doi.org/10.4111/icu.20240015

Petrosyan, A. (2024). Age distribution of internet users worldwide 2024. Statista. https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/

Somers, M. (2020). Deepfakes, explained. MIT Sloan School of Management. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Sumsub. (2023). Sumsub Research: Global Deepfake Incidents Surge Tenfold from 2022 to 2023. Sumsub. https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps. Journal of Computer-Mediated Communication, 26(6), 301–319. https://doi.org/10.1093/jcmc/zmab010

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9(11): 40-53. http://doi.org/10.22215/timreview/1282

Zeng, R., Song, S., Guo, Z., Goh, D.H., & Lee, C.S. (2023). Real or Fake: Eliciting Deepfake Identification Strategies Through a Diary Study. Proceedings of the Association for Information Science and Technology, 60(1), 1206-1208. https://doi.org/10.1002/pra2.993

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Decoding how the older adults identify misinformation: an analysis of internal mechanisms

Wu Zhaotong, Wang Yepei, and Yan Hui

DOI: https://doi.org/10.47989/ir30iConf47206

Abstract

Introduction. This study aims to explore the key factors influencing misinformation recognition among older adults, understand the internal mechanisms of how they identify misinformation, and ultimately promote age-friendly improvements in the information environment of the AI era.

Method. Using in-depth interviews, we selected 55 misinformation recognition events as research subjects and applied content analysis and binary logistic regression methods.

Analysis. Quantitative content analysis was conducted on interview data, followed by binary logistic regression analysis on nine variables, which were later consolidated into three dimensions.

Results. The findings reveal that public information sources, interpersonal information sources, information content, information perspectives, and evidence significantly impact misinformation recognition. Information sources play a crucial role in older adults' misinformation recognition, while the influence of information dimensions is not fully realized, and the direct impact of the information users' dimension is limited.

Conclusion. These findings reveal the mechanisms of misinformation identification by older adults, emphasizing the need for age-friendly improvements in information literacy and digital environments to enhance older adults' ability to recognize misinformation.

Introduction

The development of generative artificial intelligence (Generative AI), especially large language models (LLMs), has profoundly advanced transformations across various fields. '*It can creatively generate new content based on input*' (Zhang et al., 2023), significantly changing '*the production, circulation, and utilization of information and knowledge*' (Zhao & Cao, 2023). However, the powerful generative capabilities of large language models have also made the creation of misinformation more convenient and rapid. '*AI-generated content (AIGC) possesses characteristics such as complex corpus sources, variable modalities, and a combination of reality and fiction*' (Li et al., 2023), further exacerbating issues like deepfakes and expanding their scope of harm. In the era of AI algorithms, the total amount of misinformation is rapidly increasing and becoming more difficult to distinguish, leading to unprecedented harm to information users.

Among these issues, the harm of misinformation to the elderly population is particularly prominent. Due to various factors such as physical function (Wu, 2022), cognitive ability (Eich et al., 2014), digital literacy (Song & Peng, 2024), and psychological factors (Lichternberg et al., 2016), older adults are more susceptible to adopting misinformation and being misled by it, making them the primary vulnerable group to misinformation. Therefore, this study recruits older adults with experiences in recognizing and adopting misinformation, selecting misinformation recognition events obtained through in-depth interviews as research subjects. We aim to explore the main factors influencing misinformation recognition among older adults to gain a deeper understanding of the internal mechanisms of how they identify misinformation and further promote age-friendly improvements in the information environment of the AI era.

Literature review

Unlike factual information, misinformation refers to information that is false or inaccurate (Wardle & Derakhshan, 2017), often capable of intentionally or unintentionally misleading the decision-making behavior of information recipients. Against the backdrop of exponential growth in misinformation, issues related to older adults and misinformation have gained attention in academia. Existing research has compared older adults with other groups, exploring the characteristics of the elderly when facing misinformation in terms of focusing on contradictory details (Huff & Umanath, 2018), dissemination and sharing (Zhou et al., 2023), and information management (Zhou et al., 2023). Studies have focused on the relationship between older adults and various categories of misinformation such as rumours (Wu et al., 2019; Song et al., 2020), health misinformation (Cao et al., 2022; Choudhury, 2021; Peng et al., 2024), and false memories (West & Stone, 2014), with particular emphasis on the harm of misinformation to older adults (Chen, 2020; Tessoulin et al., 2020; Zhang & Su, 2019) and corresponding governance measures (Wu, 2022). However, existing research has paid limited attention to the issue of misinformation identification by older adults. Understanding the factors influencing misinformation identification among older adults is an important foundation for comprehending its internal mechanisms and has significant implications for mitigating the harm of misinformation to the elderly population. Therefore, this paper attempts to explore the internal mechanisms of misinformation identification among older adults, specifically addressing the following research questions:

> RQ1: What are the main factors older adults rely on to judge the authenticity of information?
>
> RQ2: How do these factors influence older adults' information judgment?

Research design

As part of a larger research project, this study conducts quantitative analysis based on preliminary grounded theory findings. The data were collected through in-depth interviews with 19 adults aged 60 and above who had experience in identifying and adopting misinformation. The sampling

process continued until theoretical saturation was achieved, with no new themes emerging from the final three interviews. Each 30-70 minutes interview was recorded with informed consent and transcribed, yielding over 152,000 words of standardized text materials. Interview duration varied based on participants' number of relevant experiences and communication styles, while ensuring core questions were thoroughly addressed.

The misinformation recognition events used in this study were collected from the interview texts. Initially, 83 potential cases were identified from 19 interviewees' narratives. Through fact-checking using authoritative sources, official documents, and expert consultation, we verified the authenticity of each case. After excluding cases where the information's authenticity could not be conclusively determined, 55 confirmed misinformation recognition cases from 18 elderly individuals (A1-A18) were retained for analysis. These cases covered nine major categories including politics, society, entertainment, life, health, work, finance, and knowledge. The basic characteristics of the interviewees are shown in Table 1.

Categories	Subcategories	Quantity	Categories	Subcategories	Quantity
Gender	Male	9	Education level	Elementary and below	2
	Female	9		Junior high	4
Age group	60-69 Years	4		High school/vocational	5
	70-79 Years	4		Bachelor's/associate	6
	80 and above	10		Above bachelor's	1

Table 1. Basic information of interviewees

Based on the preliminary grounded theory research, this study uses quantitative content analysis to quantify the interview data, extracting various factors relied upon by older adults in misinformation identification events. Each event case was scored dichotomously (present = 1, absent = 0) for nine variables. The coding framework and examples are shown in Table 2.

No.	Variable Name	Coding Range	Original Text Example	Variable Category
1	Public Information Sources	Focus on the credibility of public media information sources such as TV, newspapers, internet, etc. when judging the authenticity of information	TV is reliable, this information is from TV, which means it must be true to some extent	Information Source
2	Interpersonal Information Sources	Use direct interpersonal information sources as a basis for judging the authenticity of information, or be influenced by interpersonal information sources in judging the authenticity of information	It's from someone with whom I have a close relationship, won't doubt it	
3	Information Content	Judge the authenticity of information based on the words, logical relationships, and meanings of the information itself	The title and content are inconsistent, it should be untrue	Information
4	Information Viewpoints	Judge the authenticity of information based on the degree of overlap between the information viewpoint and one's own viewpoint	Content that aligns with my view, even if some details are not true, I might doubt specific details, but will generally accept it	
5	Evidence	Judge the authenticity of information based on the completeness and credibility of evidence provided in the information	It's shown in the picture, there's evidence	
6	Information Usefulness	Judge the authenticity of information based on its relevance and usefulness to oneself	Disaster information from other provinces and cities isn't relevant to me, generally I'll assume it's true	
7	Corroborating Information	Judge the authenticity of information by obtaining other information	This content (on the phone) wasn't broadcast on TV, I don't believe it	
8	Knowledge Structure	Judge the authenticity of information using existing knowledge, common sense, etc.	Older people can judge many things using common sense	Information User Factors
9	Practice	Judge the authenticity of information through personal practice after receiving the information	For health-related information, I'll try it myself and judge whether it's true through practice	

Table 2. Text coding framework and examples

Following this coding framework and standards, each of the 55 events of different categories was scored, and the statistical frequencies of each variable are shown in Table 3. Since the dependent variable (misinformation identification result) in this study is a binary variable (correct recognition = 1, incorrect recognition = 0), this paper uses IBM SPSS Statistics 26.0 to analyse the influence of independent variables through a binary logistic regression model. The data analysis is conducted in two steps: First, analysing the influence of nine original independent variables on the identification result; Second, combining the independent variables by category (see Table 2) to analyse the influence of information sources, information itself, and information users on the dependent variable.

Category	Public Info Sources	Interpersonal Info Sources	Info Content	Info Viewpoints	Evidence	Info Usefulness	Corroborating Info	Knowledge Structure	Practice	Correct Identification
Frequency	26	14	14	5	7	16	18	35	7	38

Table 3. Frequency of variable statistics

Logistic analysis of misinformation identification events among older adults

When conducting logistic analysis on the nine variables, the model accuracy was 0.73, indicating relatively high predictive accuracy. The Hosmer-Lemeshow test result ($p = 0.481 > 0.05$) showed good fit. The regression results are shown in Table 4, where Exp(B) represents the odds ratio for each variable. An odds ratio greater than 1 indicates that the variable increases the likelihood of correctly identifying misinformation, while less than 1 decreases the likelihood. The larger the value, the greater the influence of that variable.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)
Public Info Sources	1.980	0.635	9.713	1	0.002	7.245	2.085 - 25.171
Interpersonal Info Sources	0.970	0.781	1.542	1	0.215	2.637	0.571 - 12.185
Info Content	1.554	0.694	5.015	1	0.025	4.731	1.215 - 18.425
Info Viewpoints	0.894	0.807	1.227	1	0.268	2.445	0.503 - 11.887
Evidence	0.778	0.753	1.068	1	0.301	2.177	0.498 - 9.516
Info Usefulness	0.517	0.747	0.479	1	0.488	1.677	0.388 - 7.250
Corroborating Info	-0.237	0.750	0.100	1	0.753	0.789	0.182 - 3.426
Knowledge Structure	0.312	0.752	0.172	1	0.678	1.367	0.313 - 5.967
Practice	-1.820	1.037	3.078	1	0.080	0.162	0.021 - 1.237

Table 4. Logistic regression analysis results of nine variables in misinformation identification events among older adults

To further explore the relationships between variables, this study combined the 9 original independent variables into 3 dimensions:

x1 (information source) = Public information sources + Interpersonal information sources

x2 (information) = Information content + Information viewpoints + Evidence + Information usefulness + Corroborating information

x3 (information user factors) = Knowledge structure + Practice

After dimension combination, the logistic regression model accuracy was 0.82, and the Hosmer-Lemeshow test result ($p = 0.884 > 0.05$), showing significantly improved model accuracy and fit. The regression results are shown in Table 5.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)
x1	1.761	0.538	10.722	1	0.001	5.821	2.030 - 16.692
x2	1.576	0.512	9.475	1	0.002	4.838	1.771 - 13.213
x3	-0.005	0.755	0.000	1	0.995	0.995	0.227 - 4.366

Table 5. Logistic regression analysis results of three dimensions in misinformation identification events among older adults

Research findings

The logistic model analysis indicates that older adults' misinformation identification is influenced by multidimensional factors. Among these, public information sources, information content, and interpersonal information sources are the three most significant factors.

The important role of information sources in misinformation identification by older adults

Preliminary grounded theory analysis shows that older adults often rely on information sources to distinguish the authenticity of information. Most older adults consider television and newspapers as '*official channels* (A7)' when identifying information, believing that the information they publish has higher credibility, while information from mobile phones '*has a lot of false content* (A14)' and lower credibility. Moreover, older adults' identification of misinformation is also influenced by strong ties in interpersonal information sources, with some respondents stating that '*information from closer relationships is more reliable* (A9)'. In specific misinformation identification events, information from credible public information sources and strong-tie interpersonal sources is easily identified by older adults as true information, while information from less credible sources such as weak-tie interpersonal information sources undergoes more rigorous examination by older adults.

Among the nine original independent variables, focusing on public information sources ($p = 0.002$, $Exp(B) = 7.245$) is the most significant factor influencing older adults' misinformation identification. When older adults use public information sources to judge the authenticity of information, the possibility of correctly identifying misinformation is 7.245 times that of not using public information sources. The regression analysis of the three dimensions similarly shows the significance of the influence of information sources ($p = 0.001$, $Exp(B) = 5.821$). According to the frequency statistics of independent variables, in 55 events, older adults more often relied on public information sources (26 times) and interpersonal information sources (14 times) to identify misinformation. Therefore, information sources are factors that receive attention from older adults and have a significant impact on misinformation identification.

The influence of information dimensions on misinformation identification has not been fully realized

The analysis results show that the information dimension ($p = 0.002$, $Exp(B) = 4.838$) has a significant impact on older adults' identification of misinformation, with attention to the information dimension increasing the accuracy of identification by 4.838 times. Among these, information content, information viewpoints, evidence, and information usefulness all play relatively significant roles, while the influence of corroborating information is limited. Respondents use factors such as '*unrealistic details* (A11)', '*biased viewpoints* (A2)', and '*lack of evidence* (A14)' to complete their judgment on the authenticity of information. However, when identifying misinformation, older adults have limited focus on elements such as information content (14 times), viewpoints (5 times), and evidence (7 times), indicating that the influence of the information dimension on misinformation identification has not been fully realized. To further enhance older adults' ability to identify misinformation and reduce its harm, it is necessary to strengthen their ability to analyse various attributes of the information dimension and increase their attention to the authenticity of the information itself.

Limited direct influence of information user factors on misinformation identification

The regression analysis results show that the user-related factors ($p = 0.995$, $Exp(B) = 0.995$) has limited influence on older adults' misinformation identification. Combined with the previous grounded theory research findings, personal factors of older adults often indirectly influence

misinformation identification by affecting information sources and information dimensions. For example, older adults' judgment of the credibility of information sources and various attributes of the information dimension often relies on elements of knowledge structure such as industry background (A14), experience (A9), and common sense (A8), which in turn affects their identification of misinformation. Moreover, information users' factors vary greatly between individuals; older adults with more comprehensive knowledge structures can use this factor to promote misinformation identification, while those with weaker knowledge structures may be negatively inhibited by this factor.

Additionally, practice ($p = 0.080$, $Exp(B) = 0.162$) has a negative impact on the results of misinformation identification. At the same time, grounded theory research finds that when older adults verify information authenticity through practice, they have already adopted the information, and the false content within it has already caused harm to them.

Conclusion

Based on text data obtained from in-depth interviews and building upon preliminary grounded theory research, this paper conducted content analysis and binary logistic regression analysis to reveal key factors influencing misinformation identification among older adults. The results show that public information sources, interpersonal information sources, information content, information viewpoints, and evidence significantly influence misinformation identification. Furthermore, in older adults' misinformation identification, information sources play a crucial role, the influence of information dimensions has not been fully realized, and the direct influence of the information user factors is limited. These findings are significant for understanding and addressing the challenges older adults face with misinformation and are expected to provide insights for promoting age-friendly improvements in the information society.

Moreover, based on the findings of this study, we are currently examining the factors influencing misinformation identification by LLMs and their differences from older adults. In the next phase of our research, we will discuss how to better leverage LLMs to assist older adults in identifying misinformation.

Acknowledgements

I would like to express my sincere gratitude to all the elderly participants who took part in the interviews. This work was supported by the Major Project of the National Social Science Fund of China *'research on the disruptive impact of artificial intelligence on national economy and information governance'* (Grant No. 23AZD093).

About the authors

Wu Zhaotong is a master's student in Information Science at the School of Information Resource Management, Renmin University of China. Her research interests include information behavior, community informatics, and disinformation. She can be contacted at wztzndx@163.com.

Wang Yepei is an undergraduate student at the School of Information Resource Management, Renmin University of China. His research interests focus on community informatics. He can be contacted at 2022201230@ruc.edu.cn.

Yan Hui is Professor and Doctoral supervisor at the School of Information Resource Management, Renmin University of China. His research interests include community informatics and digital inequality. He can be contacted at hyanpku@ruc.edu.cn.

References

Cao, P., Liang, Y., & Liu, R. (2022). Research on public's willingness to adopt corrective health information in the context of infodemic — Based on SEM and fsQCA methods. Modern Intelligence, 42(12), 83-95.

Chen, Y. (2020). Research on rumor dissemination and audience contact during major emergencies. Journal of Shanghai Normal University (Philosophy & Social Sciences Edition), 49(6), 100-111.

Choudhury, N. (2021). Ageing, health misinformation, and mobile messaging apps. Catalan Journal of Communication & Cultural Studies, 13, 233-247. https://doi.org/10.1386/cjcs_00051_1

Eich, T. S., Murayama, K., Castel, A. D., & Knowlton, B. J. (2014). The dynamic effects of age-related stereotype threat on explicit and implicit memory performance in older adults. social Cognition, 32(6), 559-570.

Huff, M. J., & Umanath, S. (2018). Evaluating suggestibility to additive and contradictory misinformation following explicit error detection in younger and older adults. journal of Experimental Psychology: Applied, 24(2), 180-195. https://doi.org/10.1037/xap0000138

Li, B., Bai, Y., Zhan, X., & Li, G. (2023). Technical characteristics and morphological evolution of artificial intelligence generated content (AIGC). Library, Information and Knowledge, (01), 66-74. https://doi.org/10.13366/j.dik.2023.01.066

Lichternberg, P. A., Sugarman, M., Paulson, D., Ficker, L. J., & Rahman-Filipiak, A. (2016). Psychological and functional vulnerability predicts fraud cases in older adults: results of a nationally representative longitudinal study. Clinical Gerontologist, 39(2), 109-119.

Peng, W., Meng, J., & Issaka, B. (2024). Navigating persuasive strategies in online health misinformation: an interview study with older adults on misinformation management. PloS One, 19(7), e0307771. https://doi.org/10.1371/journal.pone.0307771

Song, X., Zhao, Y., Song, S., & Zhu, Q. (2020). Research on influencing factors of health rumor sharing willingness based on MOA theory. Journal of the China Society for Scientific and Technical Information, (05), 511-520.

Song, Y., & Peng, K. (2024). Digital social governance for an aging society: Reality, evolution and future approaches. Jiangxi Social Sciences, 44(5), 118-128.

Tessoulin, M., Galharret, J. M., Gilet, A. L., & Colombel, F. (2020). Misinformation effect in aging: a new light with equivalence testing. The Journals of Gerontology: Series B, 75(1), 96-103. https://doi.org/10.1093/geronb/gbz057

Vivion, M., Reid, V., Dubé, E., Coutant, A., Benoit, A., & Tourigny, A. (2024). How older adults manage misinformation and information overload - A qualitative study. BMC Public Health, 24(1), 871. https://doi.org/10.1186/s12889-024-18335-x

Wardle, C., & Derakhshan, H. (2017). Information disorder: towards an interdisciplinary framework for research and policy making. council of Europe Report.

West, R. L., & Stone, K. R. (2014). Age differences in eyewitness memory for a realistic event. The Journals of Gerontology: Series B, 69(3), 338-347. https://doi.org/10.1093/geronb/gbt014

Wu, J. (2022). Research on the spread management of health rumors among vulnerable elderly groups under public health emergencies. Scientific Research on Aging, 10(4), 66-77.

Wu, J., Sun, Y., & Li, H. (2019). Research on the spread of rumors among vulnerable elderly groups: Information, channels and audiences. Journal of Qingdao University of Science and Technology (Social Sciences), 35(3), 81-87.

Zhang, X., Yang, X., & Xu, C. (2023). Current status and future development direction of ChatGPT and generative artificial intelligence. China Science Foundation, (05), 743-750. https://doi.org/10.16262/j.cnki.1000-8217.20231026.002

Zhang, Y., & Su, Y. (2019). Research on audience identification ability of health rumors on WeChat - Based on user individual differences. audio-Visual, (10), 228- 230.

Zhao, B., & Cao, S. (2023). Test analysis of domestic and foreign generative AI large models performing typical tasks in the intelligence field. Intelligence Work, (05), 6-17.

Zhou, J., Xiang, H., & Xie, B. (2023). Better safe than sorry: A study on older adults' credibility judgments and spreading of health misinformation. Universal Access in the Information Society, 22, 957-966. https://doi.org/10.1007/s10209-022-00899-3

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Examining privacy concerns and conversations before, during, and after the onset of the COVID-19 pandemic: an analysis of r/privacy

Erica Shusas, Shruti Sannon, Katie Teitelbaum, Patrick Skeba, Eric P. S. Baumer, and Andrea Forte

DOI: https://doi.org/10.47989/ir30iConf47200

Abstract

Introduction. The COVID-19 pandemic brought about new and far-ranging uses of technology that have engendered numerous privacy concerns among the public. This study examines discussions in r/privacy, a privacy-focused online community, to gain insight into how the pandemic influenced privacy concerns and behaviours among privacy-interested people.

Method. Our research design included a mixed-method comparative analysis of 540 coded r/privacy posts from three time periods: immediately before the pandemic, during the emergence of COVID-19, and two years after the onset of the pandemic.

Analysis. We performed an inductive, qualitative analysis to understand themes in the data, as well as a statistical analysis to identify key trends and differences in the frequency of themes in the three time periods.

Results. We observed that while some changes were temporary (e.g., viewing the government as a key privacy threat during the pandemic onset), others were long-lasting and increased over time (e.g., seeking privacy-related support).

Conclusion. Our work contributes to the growing literature on how people use social media to collectively make sense of major events by exploring online discourse about privacy issues and offers insights to designers and regulators of new technologies on specific privacy concerns associated with a period of massive sociotechnical upheaval.

Introduction

The COVID-19 pandemic disrupted nearly every aspect of human life on the planet. The ways people worked (Bick et al., 2020), played (Barr & Copeland-Stewart, 2022), socialized and worshiped (Nurain et al., 2021), learned (Parthasarathy & Murugesan, 2020), exercised (Guo & Fussell, 2022), voted (Yoder et al., 2021), sought healthcare (Iacobucci, 2020), and generally lived their lives not only changed but suddenly became subject to new and unfamiliar forms of regulation and surveillance (Aloisi & De Stefano, 2022).

Both scientific literature and popular press noted dangers and actual privacy violations that occurred during the pandemic. For example, contact tracing apps, which use wireless signals to detect when one has been near a person who tested positive for COVID-19, generated scholarly (Bengio et al., 2020) and popular concern (Hautala, 2020). These concerns were validated in part by reports of Google (Ng, 2021) and of governments (Greenberg, 2020) leaking potentially identifiable information from COVID-19 apps used by millions of people. In addition to technologies specifically designed to tackle COVID-19 problems, a number of privacy violations stemmed from people's more extensive use of digital technologies. For example, the shift to videoconferencing software for work meetings and school classes led to multiple privacy concerns, from unwanted guests 'zoom-bombing' or disrupting conferences, to controversial and undisclosed data-sharing with third parties (Cox, 2020). Many people's homes became their workplaces, classrooms, and social venues, allowing others to peer inside what traditionally have been considered private spaces. On top of these technology-driven privacy issues, the social tension surrounding COVID-19 policies and vaccine statuses caused confusion and anger over how and when private health information should be disclosed and even inspired harassment and doxing attacks on public figures and other social media users (Mello et al., 2020; Yeager, 2022).

While studies have looked at people's perceptions of specific technologies during the pandemic (Häring et al., 2021; Herbert et al., 2022; Lan et al., 2020), less is known about changes in people's views on and discourse around privacy issues more broadly. Nearly five years after the outbreak of COVID most people have returned to some semblance of pre-pandemic routines and activities. However, some of the rapid adaptations we made to cope with circumstances during the height of the pandemic may have long-term societal impacts (Willige, 2021). We ask two interrelated research questions:

> **RQ1:** how did the emergence of the COVID-19 pandemic affect privacy perceptions and concerns?
>
> **RQ2:** how did online discourse around privacy shift, both during the onset of the pandemic and significantly later in the pandemic?

To answer these questions, we designed a study to understand shifts in online privacy discourse during and after the emergence of COVID-19. We collected and analysed posts on the Reddit subcommunity r/privacy from before the pandemic, during its onset, and two years later. Our approach draws heavily on ideas from crisis informatics (Palen & Anderson, 2016), particularly sensemaking, which relies on the concept 'that reality is an ongoing accomplishment that emerges from efforts to create order and make retrospective sense of what occurs' (Weick, 1993). Our paper has two primary contributions. First, our findings and analysis contribute insights into how the pandemic influenced people's privacy concerns to those designing technologies and crafting public policies on how to manage crisis events while mitigating privacy concerns. Second, we contribute to the growing body of scholarship on how people use social media to collectively make sense of crisis events by using a longitudinal approach to examine how privacy discussions evolved before, during the onset, and two years into the pandemic.

Related work

Pew Research Center reported that the majority of Americans felt that the internet was 'essential' or 'important' to them during the pandemic (McClain et al., 2021), and factors such as social distancing and shutdowns were strongly correlated to the jump in social media use seen after the onset of COVID-19 (Statista Research Department, 2022). People turned to social media for information about COVID-19 (Cinelli et al., 2020; Goel & Gupta, 2020) and for support on how to get through remote learning (Asghar et al., 2021; Huddart et al., 2020), working from home (Xiong et al., 2021) , and other challenges brought on by lockdowns (Greyling et al., 2021) and social distancing (Geirdal et al., 2021; Wiederhold, 2020). The wide-ranging use of online platforms during the pandemic raises the question of how privacy concerns have been affected and also situates online platforms as an interesting resource for understanding people's privacy-related discussions and concerns.

Crisis informatics is a multidisciplinary field that examines how people use information and communication technologies to manage concerns and uncertainties brought about by disaster events (Palen & Anderson, 2016). Concerns about the use of communication technologies are threaded throughout the lifecycle of an emergency event, from preparation to detection and recovery, and the calibre of collaboration between government, private-sector, volunteer, and citizen respondents greatly impacts resilience to crises (Pipek et al., 2014). In their review of social media within crisis informatics research, Reuter et al. note that the majority of citizen-generated social media content is meant for other citizens and that social media platforms enable citizens to assist one another during crises through the curation of 'self-help communities' (Reuter et al., 2018). Additionally, people use social media as a tool to make sense of the world around them, particularly during extreme events (Heverin & Zach, 2012; Mirbabaie et al., 2021; Stieglitz et al., 2018).

Sensemaking has frequently been used in crisis informatics to investigate how groups make sense of catastrophic events (Stieglitz et al., 2018) including campus shootings (Heverin & Zach, 2012), the Zika epidemic (Kou et al., 2017), Hurricane Irma (Chauhan & Hughes, 2020), and risks associated with COVID-19 (Pine et al., 2021). Sensemaking has also been used as a framework to explore how people make sense of privacy issues. For instance, Van Kleek et al. (Van Kleek et al., 2018) argued that the demand for non-expert users to make expert-like privacy preference decisions in complex information environments should be seen as an exercise in sensemaking. Moreover, Song et al. found that collective privacy sensemaking on Reddit helped individuals grapple with the legal and privacy vulnerabilities associated with period and fertility tracking apps after the U.S. Supreme Court overturned Roe v. Wade resulting in abortion bans in 14 states (Q. Song et al., 2024). The scope and speed of the spread of COVID-19 suggest that people turned to online discussions and spaces to make sense of the crisis. Reddit, in particular, has been investigated in relation to multiple phenomena related to COVID-19, including the collective response to media coverage of the pandemic (Gozzi et al., 2020), the variance in topical and emotional discourses related to COVID-19 among genders (Aggarwal et al., 2020), and public sentiment toward COVID-19 vaccines (Melton et al., 2021). Our study contributes to the expanding body of literature about how people collectively use social media to make sense of crisis events by exploring the changes in privacy discussions before, during, and after the onset of the COVID-19 pandemic.

Method

To understand what kinds of privacy-related threats people discussed before, during, and after the emergence of the COVID-19 pandemic, we turned to the subcommunity, r/privacy, on the popular social media platform Reddit. This is an online space where people concerned with privacy-related issues congregate to share and discuss ideas. We used the Pushshift open API (Baumgartner et al., 2020) to query Reddit data from November 2019 through December 2022. To analyze changes over time, we partitioned the data into two nine-month waves. The first was a nine-month period from

November 2019-July 2020, which included the four months before the World Health Organization's (WHO's) recognition in March 2020 of a global pandemic (November 2019–February 2020), and the five months including and following WHO's recognition (March 2020–July 2020) and totalled 23,405 posts. To determine if the changes we observed at the onset of the pandemic persisted, we then analysed data from an additional nine-month time period two years later (November 2021–July 2022), which yielded 16,092 additional posts and metadata for a total of 39,497 posts. The decision was made to select the same nine-month period two years later to control for seasonal variations.

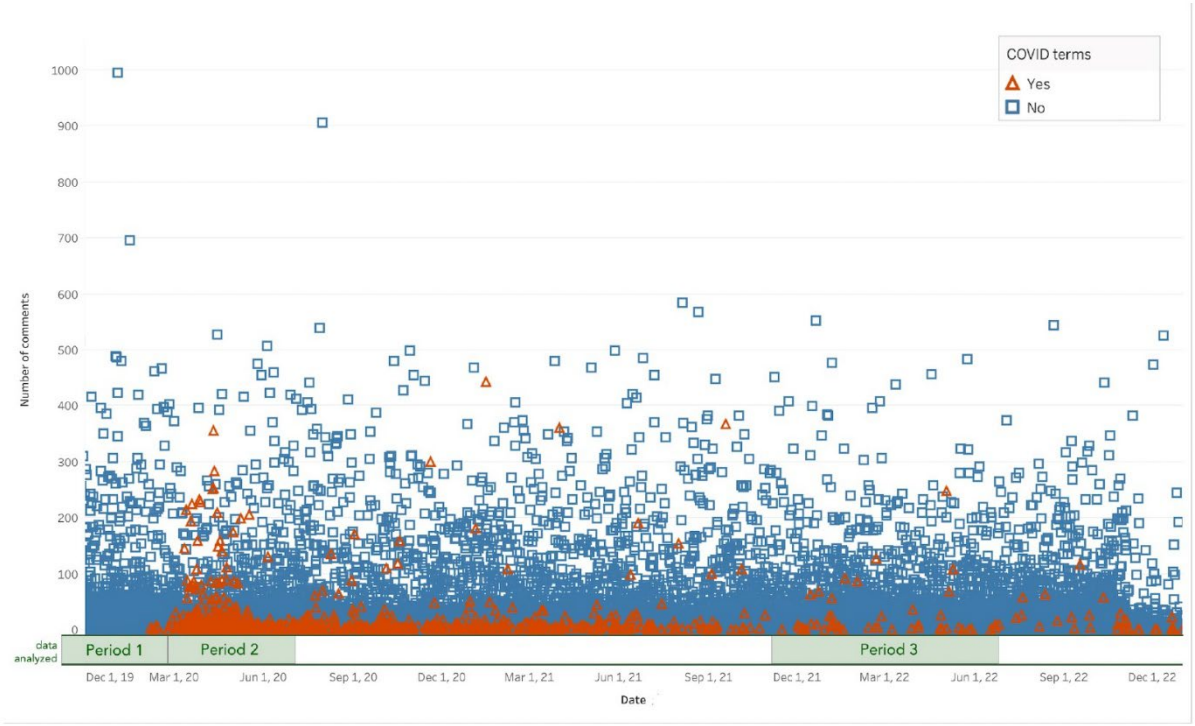


Figure 1. r/privacy posts from November 2019 through December 2022 by date posted (X-axis) and number of comments on each post (Y-axis). Blue square points indicate that the post did not contain a COVID-related term, red triangular points indicate the post contained a COVID-related term.

We used the number of comments that each post received as a proxy for the level of community attention to the topic. Comments indicate the posts to which people pay enough attention to respond; they are an internally consistent metric in that Reddit uses them (along with upvotes and other measures) to determine which posts to foreground (i.e., which posts will subsequently receive more attention). As an initial pass to identify COVID-related posts in our dataset, we chose a list of COVID-related terms (‘COVID,’ ‘*coronavirus*,’ ‘*quarantine*,’ *and* ‘*pandemic*’) and plotted the total number of posts across time to see how many contained these terms. These results (Figure 1) bolstered our choice of March 1, 2020, as a reasonable division between period 1 and period 2. To investigate changes in privacy topics on r/privacy after the emergence of COVID, we filtered the top 30 most-commented posts per month from all three data periods regardless of whether they included COVID-related terms. This yielded 540 posts. In the few instances where one of the top commented posts had been deleted, we included the next most commented-on post. Throughout the paper, we refer to these groupings of data as follows:

Period 1: Before Pandemic Onset (November 2019–February 2020)

Period 2: During Pandemic Onset (March 2020-July 2020)

Period 3: Later-stage Pandemic (November 2021-July 2022)

Analysis

We performed an inductive, qualitative analysis to understand themes in the data, as well as a statistical analysis to identify key trends and differences in the frequency of themes before and during the onset of the pandemic and whether those trends persisted two years later. The first author led all analyses. For the qualitative analysis, the first two authors initially explored data from the first and second time periods independently, using open codes and memos to identify emerging concepts. We used these concepts to build a draft codebook, on which the team iterated by doing multiple rounds of exploratory coding. The team met regularly to discuss the codebook and make necessary amendments. Categories in the finalized codebook included (1) Motivations (what the poster was seeking by posting in the forum based either on inference or directly stated by the poster), (2) Issue/Topic (what the post was about), and (3) Source (the source of the privacy threat discussed in the post).

To test inter-coder reliability, the first two authors coded small subsets of data and calculated IRR using Krippendorff's Alpha until good agreement (0.712) was reached. The first two authors then divided and coded the posts from the first and second periods containing the 30 top-commented posts from each of the 9 months for a total of 270 posts. Post title, content, and the text from URLs included anywhere in the post were included in the analysis (linked content outside of the post was not analysed). It was agreed that posts that were unrelated to privacy or too sparse to be coded would be dropped. One post was determined to be uncodeable and dropped from the analysis. In total, we analysed 269 posts from the first and second time periods.

The first and third authors did multiple rounds of coding using a subset of posts from the third time period and calculated IRR using Krippendorff's Alpha until satisfactory agreement was reached (0.903). The first and third authors then split and coded the posts from the third time period resulting in 270 posts.

We also performed a statistical analysis to identify the types of posts that changed most significantly before and after the emergence of COVID-19. An R script was used to run a multinomial logistic regression analysis. Multinomial logistic regression was chosen because it provides a single statistical analysis to ascertain which differences in proportions across the three time periods are statistically significant, as opposed to, say, conducting multiple chi-square tests. Since only a single model is constructed, all relationships (between each of the codes and the time period for a post) are tested simultaneously. Furthermore, since all codes are used as predictors, the model accounts for the simultaneous relationships each different code might have with the time period. Finally, the results of this approach can be interpreted as indicating whether a post with certain values for each code is significantly more or less likely to have been made during one of the three time periods in our dataset.

Additionally, we visualized our results using effects plots (Cook, 2009). Each predictor is categorical, with 6, 5, and 4 levels, respectively (due to their small representation in the dataset, the posts coded Other in the Motivations category and the posts coded Schools, Employers, and Other in the Source of Threat category were removed for the statistical analysis). One value is chosen as the reference level—we chose the most common values in the post-COVID data (Peng & MacKenzie, 2014), based on the Total column in Figure 2.

Ethical considerations

Posts on the subreddit r/privacy are publicly accessible. However, using public data can still give rise to ethical concerns, especially in terms of identifying individual Redditors. Thus, any quotes from posts are lightly paraphrased to protect the privacy of Redditors (Reagle & Gaur, 2022).

Findings

		Before Onset	During Onset	Later Stage	Total
Motivation	Seeking Support	4.20%	7.30%	14.07%	10.02%
	Giving Support	10.10%	8.70%	7.41%	8.35%
	Raising Awareness	68.90%	70.00%	59.63%	64.56%
	Venting	4.20%	2.00%	3.33%	3.35%
	Personal Journey	5.00%	4.00%	7.04%	5.75%
	Discussion Prompt	3.40%	8.00%	8.52%	7.24%
	Other	2.50%	0%	0%	0.56%
Issue/Topic	Privacy Concern	5.90%	7.30%	16.30%	11.50%
	Actual Violation	34%	24%	23.33%	25.97%
	Privacy Strategy	18.50%	22.70%	13.70%	17.25%
	Current Events	32.80%	40.70%	45.56%	41.37%
	Other	8.40%	5.30%	1.11%	3.90%
Source of Threat	Technology Companies	52.90%	50.70%	48.52%	48.98%
	Governments	13.40%	25.30%	18.52%	19.29%
	Schools	0.10%	2.70%	1.48%	1.67%
	Employers	0.10%	0.10%	1.11%	0.93%
	Broad/Unspecified	26.10%	20%	25.56%	24.12%
	Individual Bad Actor	1.70%	3.70%	4.81%	3.53%
	Other	4.20%	2%	0%	1.48%

Figure 2. Percentage of codes in each category before the onset, during the onset of, and in the later stage of the COVID-19 pandemic.

		During Onset		Later Stage	
Predictor	Code	Odds Ratio	p-value	Odds Ratio	p-value
Motivation	Discussion Prompt	3.28	0.0763*	5.74	0.0074*
	Giving Support	0.931	0.891	2.17	0.132
	Personal Journey	1.17	0.812	4.14	0.0162*
	Seeking Support	2.04	0.283	9.22	< 0.001*
	Venting	0.43	0.334	2.33	0.187
Issue / Topic	Privacy Concern	0.91	0.862	0.70	0.493
	Actual Violation	0.57	1.008	0.25	2.689
	Privacy Strategy	1.28	0.606	0.190	< 0.001*
	Other	0.54	3.435	0.04	1.719
Source of Threat	Individual Bad Actor	3.89	0.237	4.63	0.159
	Governments	2.006	0.058*	1.192	0.617
	Broad/Unspecified	0.649	0.231	1.03	0.927

Figure 3. Impacts of each code for each of the three predictors, one for each category, and one outcome variable, Time. Odds ratios indicate the probability of a post with that code being in the time period during the onset of the pandemic (centre columns) or two years after the onset of the pandemic (right columns) compared with before the pandemic. The most common values in the post-COVID data were used as reference levels (*Motivation: Raising Awareness*, *Issue/Topic: Current Events*, *Source of Threat: Technology Companies*). Codes that represented less than 2% of the total dataset were dropped from the analysis. *Two values were significant at the $p < 0.1$ level in During Onset: *Motivation: Discussion*, *Source of Threat: Governments*. Four values were significant at the $p < 0.1$ level in Later Stage: *Motivation: Discussion*, *Motivation: Personal Journeys*, *Motivation: Seeking Support*, *Issue/Topic: Strategy*.

In this section, we examine our thematic analysis, focusing on the themes that varied significantly across time periods. Figure 2 shows the coded proportions of posts from each time period. Figure 3 shows the results of the multinomial regression analysis, in terms of Odds Ratios (ORs) and statistical significance (p-values). While ORs provide statistical precision, they can sometimes make it difficult to interpret the simultaneous impact of a single predictor on all levels of the outcome variable. Thus, we supplement these model details with three effects plots [12] (Figures 4-6).

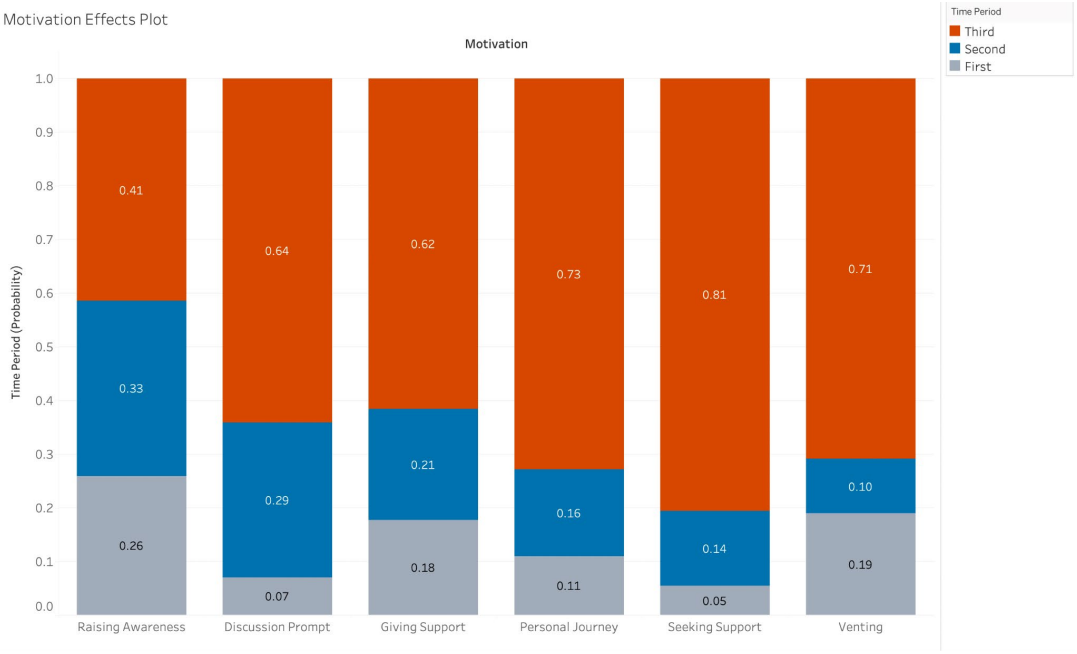


Figure 4. Visualization of Odds Ratios (ORs): Effects plot showing how different values of the Motivation code affect the probability of the post coming from each of the three time periods

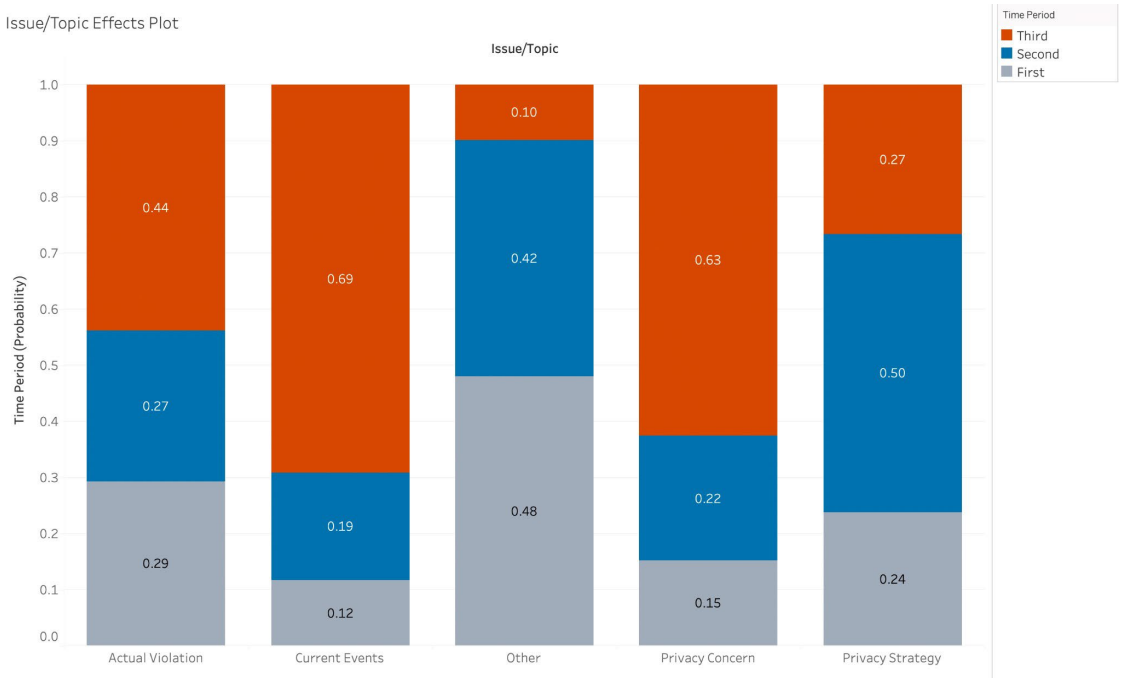


Figure 5. Visualization of Odds Ratios (ORs): Effects plot showing how different values of the Issue / Topic code affect the probability of the post coming from each of the three time periods

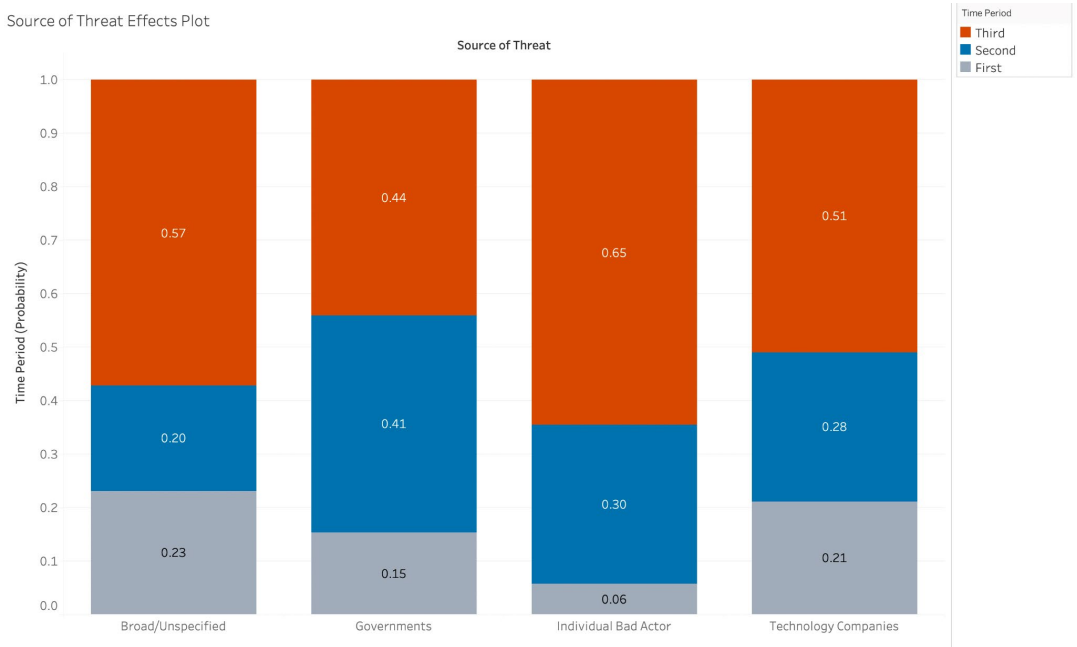


Figure 6. Visualization of Odds Ratios (ORs): Effects plot showing how different values of the Source of Threat code affect the probability of the post coming from each of the three time periods

Motivations: why post on r/privacy?

Prompting discussions

Sometimes posts served to spark a conversation about privacy issues and actively solicited feedback from the community. The number of posts containing these discussion prompts was relatively low before the pandemic, increased during the emergence of the pandemic (OR=3.28,

p=0.0763) and remained at a significantly elevated level two years later compared to before the pandemic (OR=5.74, p=0.0074). Some of the posts after the onset of the pandemic raised privacy-related perspectives unique to circumstances related to COVID-19, such as this post about face masks:

> *face masks are one of the only defences against awful facial recognition algorithms that are used against people in both public and private... How long can people continue to wear face masks? After the virus is over, will wearing face masks in public still be seen as socially acceptable? How else can we protect our privacy during the pandemic?* (Period 2 During Onset)

While others posed conceptual questions to the privacy community:

> *with all of the monstrous privacy violations that are so commonplace today, is this the end of privacy as we know it? Is there any hope for the future of privacy?* (Period 2 During Onset)

These posts were an exercise in sensemaking: they invited the community to contribute to an exchange about general privacy concerns and to discuss both personal and community stances on various privacy issues. Two years later, discussion prompts increased, for example, this query about how to safeguard privacy when living in an often-unsympathetic society:

> *Hello- I am writing this post to spark some kind of discussion about the tribulations of trying to maintain or reclaim your digital freedom and online privacy...Nobody I know cares about privacy and I'm not sure about you, but I do not have anybody that I can talk to about this IRL that wouldn't think I'm a tinfoil hat weirdo or that could actually add something to the conversation. Is this just a lost cause?* (Period 3 Later Stage)

The increase in the number of discussion prompts during the emergence of the pandemic suggests that rather than using Reddit to vent about frustrations, people turned to the community to better understand how the pandemic would impact privacy-related matters of importance to them. The persistence of higher numbers of discussion posts two years indicates that Redditors were still turning to this community to help make sense of their privacy concerns during a time of continued uncertainty.

Personal journeys

Personal journey posts also increased significantly in the later stage time period (OR=4.14, p=0.0162) (5.75%). Some posts included reflections on what shaped their current privacy practices and attitudes:

> *after paying a lot of attention to privacy news this year, I decided to stop using the same password for everything. I just spent a few hours coming up with new passwords for everything using a different strategy.* (Period 1 Before Onset)

Other posts shared struggles with negotiating their privacy preferences at work:

> *even though I told them no, my work keeps demanding I share my screen with them for pair programming. This is what I told them: "When I accepted this job, even though I agreed to work on my personal computer, I never agreed to share my screen. I find it unacceptable that this company is restricting me from what I do with my own computer. I want my personal and work life separate." I anticipate that I am probably going to have to go along with screen sharing with a virtual machine because I can't lose my job over this...* (Period 3 Later Stage)

Notably, none of the personal journey posts from our dataset refer to COVID. However, a commonality in the personal journey posts is the apparent anticipated contextual understanding

from the r/privacy community. Whether a post simply reported changing passwords or a troubling experience at work, the announcement of the personal update was rarely accompanied with an explanation of why or how the action was related to their privacy concerns–it was implied the fellow community members would understand. The increase in these types of posts suggests individuals' elevated desires in Period 3 to document life events within a supportive community.

Seeking support

These posts sought help for various privacy-related issues (10.02.%) and reflected wide-ranging concerns, from seeking informational support, such as how to prevent entities from monitoring their personal web traffic, to posts seeking emotional support, such as negotiating boundaries with family. Questions seeking support increased after the onset of the pandemic, such as this one: '*does anybody have thoughts about why all these cameras are being put up around my neighborhood*?' (Period 2 During Onset).

While we expected that people may have had more questions about keeping their privacy safe during COVID-19, we found that most seeking help posts were not looking for help around COVID-related issues. Two years later, the percentage of seeking support posts increased significantly (OR=9.22, $p < 0.001$), yet none of these posts were explicitly about pandemic-related topics. For example, this poster wanted to know if it was possible for their Facebook messages to be tracked even after they were deleted: '*if I erase my FB posts and messages will anyone still be able to find them or prove they ever existed*' (Period 3 Later Stage).

Other posts sought emotional support, such as this post on how to communicate privacy concerns to family members:

> *My mom (and dad) loves social media and have always put a lot of pictures of me and my siblings on their pages. Now that I'm older, I value my privacy and it upsets me that there are so many photos of me out there...Please let me know if I'm overreacting. I wonder if it would be rude to ask them to take the photos down* (Period 3 Later Stage).

This rise in support-seeking posts in the later stage period may be rooted in the r/privacy community experiencing elevated uncertainty about their privacy in relation to previously routine activities that were called into question after a period of rapid societal change.

Raising awareness/advocacy

The majority of posts (64.56%) throughout the period of our analysis were about raising awareness or fueling advocacy efforts on privacy topics. These efforts included posting about both potential and actual privacy issues and violations (e.g., data leaks, bad practices). Unsurprisingly, during the onset of the pandemic we also saw many news articles shared about the impact of COVID, such as: '*Covid-19 is normalising mass surveillance of citizens*' (Period 2 During Onset).

As the type of posts tended to follow the news cycle, shared headlines about COVID had petered out in Period 3 in favour of posts reflecting current events, such as: '*now that Roe v. Wade has been overturned, your abortion searches could be used to prosecute you*' (Period 3 Later Stage).

Although the topics changed, overall, we did not see a significant change in the frequency of these types of posts throughout our analysis, suggesting that these posts are an enduring and key part of r/privacy discussions. We can look at posts of this nature during different stages of the pandemic to gain insight into which issues were generating interest.

Topics: what did people post about?

Privacy strategies

Posts about privacy strategies discussed actions that could be taken to protect one's privacy. There was a bump in the proportion of posts that referred to privacy strategies at the onset of the

pandemic. We speculate that the shift from posts about actual privacy violations to privacy-protective strategies at the emergence of the pandemic might be due to community members' attempts to find ways to navigate perceived or poorly understood threats brought on by the pandemic. Some of these posts involved workarounds to COVID-related technologies:

> *iOS has contact tracing enabled by default. If you want to leave contact tracing enabled, you're automatically opted in, so you don't have to do anything, but if you want to opt out like most people, here's how you do it…* (Period 2 During Onset)

Some community members recognized that people had to use technologies that could be problematic for privacy during the pandemic due to the shift to remote work. In such cases, posts shared tips for protecting privacy while using these technologies, such as the following post about videoconferencing software: '*we're all probably tired of hearing about the Zoom privacy fiasco. If you have to use it, here's some tips to do it safer…*' (Period 2 During Onset).

Two years later, the percentage of posts about privacy protective strategies had declined significantly compared with pre-pandemic levels (OR=0.190, $p < 0.001$) and none of the posts in this category in our sample referred to COVID specifically. This decline might be a result from a downtrend in the need to rapidly adopt new strategies to adjust to incomparable circumstances brought on by the emergence of COVID.

Privacy concerns

Posts about privacy concerns discussed a worry or uncertainty about a privacy issue but did not explicitly report an actual privacy violation. The percentage of these types of posts was relatively stable from before COVID-19 through the onset of the pandemic, but there was an increase during the later stage time period. During Period 3, concerns were commonly about either technology company developments or policy changes ('*Will you continue to use DuckDuckGo after this announcement?*' (Period 3 Later Stage)) or concerns about potential violations ('*Is a Linux or Apple laptop more difficult for law enforcement to get into?*' (Period 3 Later Stage)). Overall, this shift from posts about privacy strategies to privacy concerns reveals a transition from an interest in action-oriented posts to those of suspicion and unease.

Current events

Posts about current events included discussions about new laws or court cases, public stances by politicians, or other current events in the news. These posts were frequent both before and during the pandemic, though they increased in frequency after the start of the pandemic (32.8% to 40.7%) and increased further two years later (45.56%). In the four months following the emergence of COVID-19, several posts focused on how the pandemic was impacting privacy concerns and violations. For example, the following headline considered the impact of remote education on students' privacy and safety: '*teacher Spying on Student During Virtual Class Sends Cops to Search 11-Year-Old's Home After Spotting a BB Gun*' (Period 2 During Onset).

Posts that called out COVID by name had nearly ceased by the later stage time period, but most of the posts that mentioned COVID were current event posts, such as this one: '*CDC Monitored Millions of Phones to Discover If Americans Followed COVID Lockdown Orders*' (Period 3 Later Stage).

The increase in attention to current events posts during the onset of COVID-19 indicates a heightened interest in breaking and developing news stories that continued to grow in the later stage of the pandemic.

What were the sources of perceived threats?

Governments

The greatest increase in source of threat after the emergence of the pandemic was by governments (OR=2.006, p=0.058). After COVID-19 emerged, posts that cited governments as the main source of threat often highlighted current actions (or suspected actions) involving governments. Some of these posts relayed official announcements:

> *today, Argentinian President Alberto Fernández announced that workers returning to work after lockdown will be required to download an application for their phones which will track and record their location* (Period 2 During Onset).

Other posts detailed government actions that might have been occurring covertly:

> *the FBI is promoting a home-workout app for people stuck indoors during the coronavirus outbreak. It also tracks your location and which WiFi networks you're using* (Period 2 During Onset).

In addition to specific technologies, apps, and services being espoused by governments in response to the pandemic, several posts also voiced concerns about governments using the pandemic to put broader surveillance structures in place: '*Snowden Warns Governments Are Using Coronavirus to Build 'the Architecture of Oppression*' (Period 2 During Onset).

It is interesting to note here that rather than posting about actual violations that had been reported where the government was identified as the source of threat, there was an increase in posts that questioned how governments were balancing (or taking advantage of) measures to control COVID-19 and the privacy rights of individuals. The frequency of posts with governments as a source of threat had decreased to 18.52% two years later. As the pandemic wore on and some of the restrictions lessened, so did the appearance of governments as a source of threat posts in our dataset.

Technology companies

Throughout our period of data collection, technology companies were the most frequently cited source of privacy threats (48.98%). Perhaps unsurprisingly, posts that were concerned about tech companies more frequently referenced Zoom and other video conferencing technologies after the pandemic began. Some of these posts contained links to stories that included factual information about problems with Zoom, such as the following: '*ex-NSA hacker finds new Zoom flaws to takeover Macs again, including webcam, mic, and root access*' (Period 2 During Onset).

In contrast, some posts were impassioned appeals, likely brought about by the sudden and ubiquitous use of video conferencing software in several domains of life, including education and work, such as 'For the love of God, stop using Zoom!' (Period 2 During Onset). Interestingly, the number of mentions of Zoom or other videoconferencing tools dropped back down to zero just two years later—perhaps an indication of how interwoven these technologies have become in people's daily routines.

Discussion

The impact of COVID-19 on people's privacy perceptions and concerns

During the rapid technological shifts caused by the pandemic, people became increasingly dependent on technologies—both new and existing—that were integrated into many aspects of daily life. In our data, we saw that seeking-support posts increased at the onset of the pandemic. This signals that people felt their existing privacy strategies were insufficient to protect them from threats they were experiencing at the time. Thus, at the onset of the pandemic, the subreddit became a space where posters increasingly sought ways to mitigate unfamiliar obstacles. By

identifying these trends, attention can further be drawn toward the specific issues being discussed during periods of rapid societal change. This may help those designing, implementing, and regulating new technologies and strategies to deal with evolving crises by promptly identifying privacy flaws or areas of concern. For example, at the onset and in the later stage of the pandemic we observed an uptick in support-seeking related to suspicions about phones or other devices being tracked by individual actors such as family members, neighbours, employers, or schools. Although these posts did not discuss COVID-related technologies, a reasonable interpretation is that media coverage and discussion of privacy concerns about contact tracing apps had stoked fears about the possibility of being tracked by other entities. Designing for transparency around tracking capabilities or enabling tracking preferences becomes a more crucial privacy design recommendation in a world with heightened concerns about tracking.

At the onset of the pandemic, we also saw a growth in posts about governmental actors as the source of threat. This finding is likely because when the pandemic emerged, governments were the most salient entities responsible for changes to people's lives via COVID-related mandates. The relationship between volunteer/community resources and official governmental responses is a resurfacing theme in crisis informatics research (Douglas et al., 2018; Kavanaugh et al., 2011); as per the crisis informatics community, decreasing loss of life and damages caused by crises depends on successful collaboration between governments, professionals, volunteers, and citizen respondents (Pipek et al., 2014). Prior work finds that the dissemination of quality information on social media helps increase the public's trust in government (Arshad & Khurram, 2020; Bertot et al., 2010; C. Song & Lee, 2016). Our findings suggest that during the onset of the pandemic, when there was a significant increase in posts identifying governments as a source of threat, it may have been advantageous for government agencies to increase their usage of social media as a channel to circulate information on privacy policies related to new public health technologies and practices. Governments can look to social media channels such as r/privacy during periods of crisis or rapid social change as a way to probe levels of public confidence, identify misconceptions, and work toward mitigating privacy issues that raise concerns among citizens.

Collaborative sensemaking in online discourse around privacy during the pandemic

We observed an increase in posts that sought to spur discussions about privacy-related topics on r/privacy during the onset of the pandemic. Our longitudinal approach revealed that two years later, this trend had continued. This trend suggests a sustained interest among r/privacy users in others' perspectives on privacy-related dilemmas. By turning to an online community of like-minded people, r/privacy members had opportunities to collaboratively make sense of the privacy implications of the technological shifts brought about by the pandemic. Research finds that people collaborate to make sense of privacy issues (Murthy et al., 2021; Q. Song et al., 2024; Watson et al., 2020), as well as to organize against privacy threats through collective action (Das et al., 2021; Vincent et al., 2021). Our findings suggest that, once established, the practice of inviting collaborative sensemaking persisted in r/privacy.

The data also revealed a significant increase in posts about users' personal journeys related to privacy. These posts were similar to diary or journal entries—they reported personal privacy-related activities or milestones for seemingly no other purpose than to have them on record. Diaries have been looked at as sensemaking activities (Elsden et al., 2016) and qualitative diary data has been used to offer insight to policymakers about people's perspectives, experiences, and coping mechanisms during COVID-19 (Mueller et al., 2023; Terzis et al., 2022). Our findings suggest that researchers may be able to turn to online communities to access some of these traditionally private sensemaking practices. Like, blogs, posts about personal journeys in our dataset differ from traditional diary entries in that they are publicly shared with an online community. We speculate

that self-disclosures serve to reinforce the sense of community, establish norms, and create a conducive environment for sensemaking.

Limitations

Since we analysed a subreddit that focuses on privacy issues, our findings reflect the privacy concerns and attitudes of people who care about privacy and actively seek discussions about it. We believe that focusing on this community gives us valuable insight into how a community collectively navigates privacy issues; however, our data does not reflect the impact of the pandemic on people for whom privacy is less of a salient concern or who do not choose to discuss privacy issues online.

Further, while we focused on the effect that COVID-19 had on privacy concerns, we note that the pandemic was not the only major event that took place during the time period of our analysis. For example, some of the posts in our dataset referenced privacy-related concerns in the wake of protests occurring after the murder of George Floyd in 2020. These events may also contribute to our findings, especially around governmental actors being seen as a source of threat.

Conclusion

Robert Soden has called for a long view of crises, writing, 'Disasters reverberate through the history of the places in which they occur. Long after the period of crisis is deemed to be over, they continue to have impact, foreclosing some possible futures, and opening others' (Soden & Palen, 2018). Inspired by Soden, we aimed to understand how people's privacy concerns and behaviours may have changed and may continue to change in light of an enduring crisis that has unfolded over multiple years—the COVID-19 pandemic. We conducted a thematic analysis of the 540 posts that received the highest number of comments on r/privacy to understand the most commonly discussed privacy topics, concerns, and behaviours during these time periods. Although some changes were temporary (e.g., viewing the government as a key privacy threat during the pandemic onset), others were long-lasting and increased over time (e.g., seeking privacy-related support). These findings contribute to growing literature on how people use social media to collectively make sense of catastrophic events by providing insight into privacy-related collaborative sensemaking practices during a crisis. Further, by examining privacy discussions during multiple points of the pandemic, we show that privacy concerns and behaviours in the wake of a global crisis are continuously shifting and warrant attention over time.

Acknowledgements

This work was supported by the National Science Foundation (Awards #1816264, #1814533, and #2439312, formerly #2016058). The findings and recommendations in this paper are those of the authors and do not necessarily reflect the views of the NSF.

About the authors

Erica Shusas is an Information Science PhD student at Drexel University. She holds an MA in Media Studies from The New School and a BA in Social Thought and Political Economy from the University of Massachusetts, Amherst. Her dissertation work aims to understand how young people assess the credibility of content labels for misinformation and AI-generated content on social media platforms. Her research interests include social computing, privacy, credibility assessment, and AI-generated content. She may be contacted at ejr93@drexel.edu.

Dr. Shruti 'Tee' Sannon is a digital rights advocate working in the field of technology policy. She holds a PhD in Communication from Cornell University, where she specialized in human-computer interaction and social computing. She also holds an MA in Media, Culture, and Communication

from NYU and a BS in Psychology and Criminology from the University of Toronto. Her research interests include privacy, surveillance, AI, and digital labour.

Katie Teitelbaum holds an MS in Human-computer Interaction and User Experience Design from Drexel University and an Interdisciplinary BA in User Interface / User Experience with minors in Chinese and Computer Science from Lehigh University.

Dr. Patrick Skeba is a Teaching Assistant Professor at the School of Computing and Information at the University of Pittsburgh. He received a PhD in Computer Science from Lehigh University and bachelor's degrees in Cognitive Science and Computer Science from Johns Hopkins University. His teaching and research interests include internet privacy, machine learning, and responsible use of AI and data.

Dr. Eric P. S. Baumer is an Associate Professor of Computer Science and Engineering at Lehigh University. Dr. Baumer holds an MS and PhD in Information and Computer Sciences from the University of California, Irvine, a BS in Computer Science with a minor in Music from the University of Central Florida and held a post-doctoral research position at Cornell University. His research interests include human-computer interaction, social informatics, technology resistance and non-use, and natural language processing.

Dr. Andrea Forte is Dean of the University of Michigan School of Information (UMSI). She holds a PhD in Human-Centered Computing from Georgia Institute of Technology, an MLIS in Library and Information Science from the University of Texas at Austin, and a BA in Language and Philosophy from Western Michigan University. Her research interests include HCI, computer-supported cooperative work, social computing, privacy, and computing education.

References

Aggarwal, J., Rabinovich, E., & Stevenson, S. (2020). Exploration of Gender Differences in COVID-19 Discourse on Reddit. arXiv:2008.05713 [Cs]. https://doi.org/10.48550/arXiv.2008.05713

Aloisi, A., & De Stefano, V. (2022). Essential jobs, remote work, and digital surveillance: Addressing the COVID-19 pandemic panopticon. International Labour Review, 161(2), 289–314. https://doi.org/10.1111/ilr.12219

Arshad, S., & Khurram, S. (2020). Can government's presence on social media stimulate citizens' online political participation? Investigating the influence of transparency, trust, and responsiveness. Government Information Quarterly, 37(3), 101486. https://doi.org/10.1016/j.giq.2020.101486

Asghar, M. Z., Iqbal, A., Seitamaa-Hakkarainen, P., & Barbera, E. (2021). Breaching Learners' Social Distancing through social media during the COVID-19 Pandemic. International Journal of Environmental Research and Public Health, 18(21), 11012. https://doi.org/10.3390/ijerph182111012

Barr, M., & Copeland-Stewart, A. (2022). Playing video games during the COVID-19 pandemic and effects on players' well-being. Games and Culture, 17(1), 122–139.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. Proceedings of the International AAAI Conference on Web and social media, 14, 830–839.

Bengio, Y., Janda, R., Yu, Y. W., Ippolito, D., Jarvie, M., Pilat, D., Struck, B., Krastev, S., & Sharma, A. (2020). The need for privacy with public digital contact tracing during the COVID-19 pandemic. The Lancet Digital Health, 2(7), e342–e344. https://doi.org/10.1016/S2589-7500(20)30133-3

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. Government Information Quarterly, 27(3), 264–271. https://doi.org/10.1016/j.giq.2010.03.001

Bick, A., Blandin, A., & Mertens, K. (2020). Work from home after the COVID-19 outbreak. https://doi.org/10.24149/wp2017

Chauhan, A., & Hughes, A. L. (2020). Trustworthiness Perceptions of Social Media Resources Named after a Crisis Event. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), 1–23. https://doi.org/10.1145/3392849

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. Scientific Reports, 10(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Cook, R. D. (2009). Regression graphics: Ideas for studying regressions through graphics. John Wiley & Sons.

Cox, J. (2020, March 26). Zoom iOS App Sends Data to Facebook Even if You Don't Have a Facebook Account. VICE. https://www.vice.com/en/article/zoom-ios-app-sends-data-to-facebook-even-if-you-dont-have-a-facebook-account/#:~:text=What%20the%20company%20and%20its,Motherboard%20analysis%20of%20the%20app.

Das, S., Edwards, W. K., Kennedy-Mayo, D., Swire, P., & Wu, Y. (2021). Privacy for the People? Exploring Collective Action as a Mechanism to Shift Power to Consumers in End-User Privacy. IEEE Security & Privacy, 19(5), 66–70. https://doi.org/10.1109/MSEC.2021.3093135

Douglas, G., Klinenberg, E., & Koslov, L. (2018). Grassroots Relief: Informal and Community-Based Response to Extreme Weather Events from Occupy Sandy to the Cajun Navy.

Elsden, C., Durrant, A. C., & Kirk, D. S. (2016). It's Just My History Isn't It?: Understanding Smart Journaling Practices. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2819–2831. https://doi.org/10.1145/2858036.2858103

Geirdal, A. Ø., Ruffolo, M., Leung, J., Thygesen, H., Price, D., Bonsaksen, T., & Schoultz, M. (2021). Mental health, quality of life, wellbeing, loneliness, and use of social media in a time of social distancing during the COVID-19 outbreak. A cross-country comparative study. Journal of Mental Health, 30(2), 148–155. https://doi.org/10.1080/09638237.2021.1875413

Goel, A., & Gupta, L. (2020). Social Media in the Times of COVID-19. Journal of Clinical Rheumatology: Practical Reports on Rheumatic & Musculoskeletal Diseases, 26(6), 220–223. PubMed. https://doi.org/10.1097/RHU.0000000000001508

Gozzi, N., Tizzani, M., Starnini, M., Ciulla, F., Paolotti, D., Panisson, A., & Perra, N. (2020). Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis. Journal of Medical Internet Research, 22(10), e21597. https://doi.org/10.2196/21597

Greenberg, A. (2020, May 6). India's Covid-19 Contact Tracing App Could Leak Patient Locations. Wired. https://www.wired.com/story/india-covid-19-contract-tracing-app-patient-location-privacy/

Greyling, T., Rossouw, S., & Adhikari, T. (2021). The good, the bad and the ugly of lockdowns during Covid-19. PLOS ONE, 16(1), e0245546. https://doi.org/10.1371/journal.pone.0245546

Guo, J., & Fussell, S. R. (2022). 'It's Great to Exercise Together on Zoom!': Understanding the Practices and Challenges of Live Stream Group Fitness Classes. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1), 1–28. https://doi.org/10.1145/3512918

Häring, M., Gerlitz, E., Tiefenau, C., Smith, M., Wermke, D., Fahl, S., & Acar, Y. (2021). Never ever or no matter what: Investigating Adoption Intentions and Misconceptions about the {Corona-Warn-App} in Germany. Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021), 77–98.

Hautala, L. (2020). COVID-19 contact tracing apps create privacy pitfalls around the world [Online post]. https://www.cnet.com/news/privacy/covid-contact-tracing-apps-bring-privacy-pitfalls-around-the-world/

Herbert, F., Kowalewski, M., Schnitzler, T., Lassak, L., & Dürmuth, M. (2022). "Fast,Easy,Convenient." StudyingAdoptionandPerceptionofDigitalCovidCertificates. Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022) Boston, MA, USA.

Heverin, T., & Zach, L. (2012). Use of microblogging for collective sense-making during violent crises: A study of three campus shootings. Journal of the American Society for Information Science and Technology, 63(1), 34–47. https://doi.org/10.1002/asi.21685

Huddart, D., Hirniak, J., Sethi, R., Hayer, G., Dibblin, C., Meghna Rao, B., Ehsaanuz Zaman, M., Jenkins, C., Hueso, B., & Sethi, S. (2020). #MedStudentCovid: How social media is supporting students during COVID-19. Medical Education, 54(10), 951–952. https://doi.org/10.1111/medu.14215

Iacobucci, G. (2020). Covid-19: Diabetes clinicians set up social media account to help alleviate patients' fears. BMJ, m1262. https://doi.org/10.1136/bmj.m1262

Kavanaugh, A. L., Yang, S., Li, L. T., Sheetz, S. D., & Fox, E. A. (2011). Microblogging in crisis situations: Mass protests in Iran, Tunisia, Egypt.

Kou, Y., Gui, X., Chen, Y., & Pine, K. (2017). Conspiracy Talk on Social Media: Collective Sensemaking during a Public Health Crisis. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1–21. https://doi.org/10.1145/3134696

Lan, G. Z., Zhao, Z., & Lei, L. (2020). The Use of Health Code to Control the COVID-19 Pandemic: Achievements, Evaluations and Future Prospects. 2020 5th International Conference on Universal Village (UV), 1–5. https://doi.org/10.1109/UV50937.2020.9426225

McClain, C., Vogels, E. A., Andrew Perrin, Sechopoulos, Stella, & Raine, L. (2021). The Internet and the Pandemic. Pew Research Center. https://www.pewresearch.org/internet/2021/09/01/the-internet-and-the-pandemic/

Mello, M. M., Greene, J. A., & Sharfstein, J. M. (2020). Attacks on Public Health Officials During COVID-19. JAMA, 324(8), 741. https://doi.org/10.1001/jama.2020.14423

Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. Journal of Infection and Public Health, 14(10), 1505–1512. https://doi.org/10.1016/j.jiph.2021.08.010

Mirbabaie, M., Amojo, I., & Stieglitz, S. (2021). Affording Twitter in Emergency Situations: The Occurrence of Rumor Sense-Making. Journal of Database Management, 32(2), 50–66. https://doi.org/10.4018/JDM.2021040104

Mueller, G., Barford, A., Osborne, H., Pradhan, K., Proefke, R., Shrestha, S., & Pratiwi, A. M. (2023). Disaster Diaries: Qualitative Research at a Distance. International Journal of Qualitative Methods, 22, 160940692211471. https://doi.org/10.1177/16094069221147163

Murthy, S., Bhat, K. S., Das, S., & Kumar, N. (2021). Individually Vulnerable, Collectively Safe: The Security and Privacy Practices of Households with Older Adults. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–24. https://doi.org/10.1145/3449212

Ng, A. (2021). Google Promised Its Contact Tracing App Was Completely Private—But It Wasn't. The Markup. https://themarkup.org/privacy/2021/04/27/google-promised-its-contact-tracing-app-was-completely-private-but-it-wasnt

Nurain, N., Chung, C.-F., Caldeira, C., & Connelly, K. (2021). Hugging with a Shower Curtain: Older Adults' Social Support Realities During the COVID-19 Pandemic. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1–31. https://doi.org/10.1145/3479607

Palen, L., & Anderson, K. M. (2016). Crisis informatics—New data for extraordinary times. Science, 353(6296), 224–225. https://doi.org/10.1126/science.aag2579

Parthasarathy, S., & Murugesan, S. (2020). Overnight Transformation To Online Education Due to the COVID-19 Pandemic: Lessons learned. eLearn, 2020(9), 3424971.3421471. https://doi.org/10.1145/3424971.3421471

Peng, D., & MacKenzie, G. (2014). Discrepancy and Choice of Reference Subclass in Categorical Regression Models. In G. MacKenzie & D. Peng (Eds.), Statistical Modelling in Biostatistics and Bioinformatics (pp. 159–184). Springer International Publishing. https://doi.org/10.1007/978-3-319-04579-5_12

Pine, K. H., Lee, M., Whitman, S. A., Chen, Y., & Henne, K. (2021). Making Sense of Risk Information amidst Uncertainty: Individuals' Perceived Risks Associated with the COVID-19 Pandemic. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3411764.3445051

Pipek, V., Liu, S. B., & Kerne, A. (2014). Crisis Informatics and Collaboration: A Brief Introduction. Computer Supported Cooperative Work (CSCW), 23(4–6), 339–345. https://doi.org/10.1007/s10606-014-9211-4

Reagle, J., & Gaur, M. (2022). Spinning words as disguise: Shady services for ethical research? First Monday. https://doi.org/10.5210/fm.v27i1.12350

Reuter, C., Hughes, A. L., & Kaufhold, M.-A. (2018). Social media in crisis management: An evaluation and analysis of crisis informatics research. International Journal of Human–Computer Interaction, 34(4), 280–294.

Soden, R., & Palen, L. (2018). Informating Crisis: Expanding Critical Perspectives in Crisis Informatics. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1–22. https://doi.org/10.1145/3274431

Song, C., & Lee, J. (2016). Citizens' Use of Social Media in Government, Perceived Transparency, and Trust in Government. Public Performance & Management Review, 39(2), 430–453. https://doi.org/10.1080/15309576.2015.1108798

Song, Q., Ma, R., Kou, Y., & Gui, X. (2024). Collective Privacy Sensemaking on Social Media about Period and Fertility Tracking post Roe v. Wade. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1–35. https://doi.org/10.1145/3641000

Statista Research Department,. (2022). Average daily time spent on social networks by users in the United States from 2018 to 2022. Statista. https://www.statista.com/statistics/1018324/us-users-daily-social-media-minutes/

Stieglitz, S., Bunker, D., Mirbabaie, M., & Ehnis, C. (2018). Sense-making in social media during extreme events. Journal of Contingencies and Crisis Management, 26(1), 4–15. https://doi.org/10.1111/1468-5973.12193

Terzis, L. D., Saltzman, L. Y., Logan, D. A., Blakey, J. M., & Hansel, T. C. (2022). Utilizing a Matrix Approach to Analyze Qualitative Longitudinal Research: A Case Example During the COVID-19 Pandemic. International Journal of Qualitative Methods, 21, 160940692211237. https://doi.org/10.1177/16094069221123723

Van Kleek, M., Seymour, W., Veale, M., Binns, R., & Shadbolt, N. (2018). The need for sensemaking in networked privacy and algorithmic responsibility. Sensemaking in a Senseless World: Workshop at ACM CHI'18, 22 April 2018, Montréal, Canada.

Vincent, N., Li, H., Tilly, N., Chancellor, S., & Hecht, B. (2021). Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 215–227. https://doi.org/10.1145/3442188.3445885

Watson, H., Moju-Igbene, E., Kumari, A., & Das, S. (2020). 'We Hold Each Other Accountable': Unpacking How Social Groups Approach Cybersecurity and Privacy Together. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3313831.3376605

Weick, K. E. (1993). The collapse of sensemaking in organizations: The Mann Gulch disaster. Administrative Science Quarterly, 628–652.

Wiederhold, B. K. (2020). Social Media Use During Social Distancing. Cyberpsychology, Behavior, and Social Networking, 23(5), 275–276. https://doi.org/10.1089/cyber.2020.29181.bkw

Willige, A. (2021). What are the long-term effects of the COVID-19 pandemic? These 5 trends give us a glimpse [Online post]. World Economic Forum. https://www.weforum.org/agenda/2021/08/covid19-long-term-effects-society-digital/

Xiong, Z., Li, P., Lyu, H., & Luo, J. (2021). Social Media Opinions on Working From Home in the United States During the COVID-19 Pandemic: Observational Study. JMIR Medical Informatics, 9(7), e29195. https://doi.org/10.2196/29195

Yeager, V. A. (2022). The Politicization of Public Health and the Impact on Health Officials and the Workforce: Charting a Path Forward. American Journal of Public Health, 112(5), 734–735. https://doi.org/10.2105/AJPH.2022.306744

Yoder, J., Handan-Nader, C., Myers, A., Nowacki, T., Thompson, D. M., Wu, J. A., Yorgason, C., & Hall, A. B. (2021). How did absentee voting affect the 2020 U.S. election? Science Advances, 7(52), eabk1755. https://doi.org/10.1126/sciadv.abk1755

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

'Everyone has their reasons for curating the data they have decided to keep': a thematic analysis of data hoarding as digital curation practice

Emily Maemura and Travis L. Wagner

DOI: https://doi.org/10.47989/ir30iConf47197

Abstract

Introduction. This paper presents a preliminary analysis of the r/DataHoarder subreddit, an online community focused on social, legal, and technical practices related to digital curation. We contrast how they conceptualize their practices with the models, frameworks, and capabilities established for digital curation in more traditional memory institutions.

Method. We use thematic analysis to analyse top posts (n=170) from the subreddit, to determine how members conceptualize, describe, and enact their data hoarding practices.

Findings. Two key themes are identified: a focus on materiality of storage for hoarding, and use of the subreddit to promote calls-to-action. Each theme is further analysed, identifying underlying motivations such as nostalgia and vigilantism.

Discussion. We briefly discuss the theoretical and practical implications of these findings in contrast with workflows like the DCC curation lifecycle, while also addressing limitations of this preliminary work, and outlining potential new research directions moving forward.

Conclusion. We have found this analysis presents useful counterpoints to commonly referenced and standardised practices of memory institutions. We believe this will continue to be a fruitful area of study as we conduct future work on how r/DataHoarder and similar communities conceptualise, practice, and develop their own ethical framework toward sustained access of digital materials.

Introduction

The r/DataHoarder subreddit is an online community focused on social, legal, and technical practices related to digital curation. Created in May 2013 and with a current membership of over 700,000 users, it is neither representative of any institution nor defined by a shared digital curation credo. Yet, r/DataHoarder nonetheless defines itself as a community of '*digital librarians*' who have varied '*reasons for curating the data they have decided to keep (either forever or For A Damn Long Time™)*' (r/DataHoarder, 2017). Regardless of type of data or reasons for hoarding, r/DataHoarder members share best practices, technical guidance, and, on occasion, ethical stances towards the increasingly complex sociotechnical landscape of a digital-first world. As such, r/DataHoarder acts as a distributed community of practice, enacting digital curation through deinstitutionalized and extralegal means.

A recent example of r/DataHoarder's collective efforts is seen in their archiving work in the wake of the January 6th United States Capitol attack. The subreddit's users collaborated to document and preserve footage from the attack as well as communications between those involved, which received both scholarly and journalistic attention (Basu, 2021; Chapman, 2022). This example demonstrates how, despite a distributed community structure, r/DataHoarder's members also coalesce around shared ethics and practices. We take an interest in exploring these motivations, where they both align with professional guidelines, or alternately exhibit distrust for traditional collecting institutions (i.e., government agencies), and how they work intentionally outside these institutional boundaries.

This study therefore examines how r/DataHoarder imagines and actualizes digital curation practice. We consider how such framings align with and depart from standards and practices of digital curation within traditional memory institutions. We present a preliminary thematic analysis of the top posts which have a vote count of 2,000 or more (Panek, 2021). We present this exploratory study to address the following research questions:

- What is the practice of '*data hoarding*' and what activities and capabilities does it entail?
- Where do these activities and capabilities intersect with or diverge from those of memory institutions' practices of curation?
- What do these practices reveal of the ethics and social ideologies underlying data hoarding work, and motivations of individual data-hoarders?

Literature review

Over the past three decades, scholarship in digital preservation and curation has produced a knowledge base of best practices, primarily focused on conducting this work within organizational settings. This includes the standardization of preservation systems and functions through the open archival information system reference model, as well as common practices and activities as referenced in the digital preservation coalition (DPC) handbook and outlined in the digital curation centre (DCC) lifecycle. The specific needs for organizational and individual skillsets in conducting preservation and curation have also been studied, prescribed, and measured through capability models and tools like the NDSA levels of preservation, DPC competency audit tool, and DLF's levels of born-digital access (NDSA, 2019; McMeekin & Currie, 2022; Peltzman et al., 2022).

Largely separate from these organizational approaches, we see increased public awareness of digital content at risk of loss. Recent grassroots collecting of digital materials include: work by the environmental data governance initiative to collect EPA data prior to the beginning of President Trump's term in office in 2017 (Walker et al., 2018); work by archive team to collect NSFW materials following Tumblr's policy change and removals in 2018 (Ogden, 2022); efforts of Saving Ukrainian Cultural Heritage Online (SUCHO) to preserve digital materials from Ukraine following the Russian invasion in February 2022 (LeBlanc et al., 2022). We observe that these forms of collecting, and

preservation differ significantly from the work by practitioners within traditional memory institutions. We take inspiration from Dallas (2016) who advises digital curation research should,

> *focus on the study of actual practices of curation in a diversity of contexts … and should prioritize such intellectual inquiry over the imposition of models developed within the custodial fold* (p. 440).

We therefore take up a concern with understanding collecting and curation outside of memory institutions.

For this initial, exploratory study, we frame this work as a '*community of practice*' to understand how groups of diverse and distributed individuals engage in digital curation. Coined by Lave and Wenger (1991), the concept of '*communities of practice*' (CoPs) originates in organizational studies. Cox (2005) addresses the concept's varied usage and distinctions between the early applications of CoPs, and later take-up by Brown and Duguid (1991, 2001), as well as Wenger's (1998, 2000) subsequent publications. In information studies, the concept is often deployed to describe or facilitate efforts in community building around shared practices, both within and across organizations. For digital curation in particular, CoPs have been the guiding framework for developing capacity-building workshops at the local and national level (Moran et al., 2019, Rios et al. 2020).

We adopt a broader view of CoPs, as advocated by Bowker and Star (2000) who place less emphasis on the organizational setting. This understanding of CoPs has been previously applied to study the knowledge practices of amateur and hobby communities, and particularly their distributed formation in virtual space (Hills, 2015). Several precedents also explore this type of community formation for learning and sharing practices specifically through reddit and its structure of subreddits (Britt et al., 2022; Haythornthwaite et al., 2018; Hudgins et al., 2020; Kwon et al. 2020). Proferes et al. (2021) provide a systematic overview of past research with reddit, which we also use to inform our choice in data collection, scale and methods of analysis, and ethical decisions for working with public but potentially sensitive and personal posts and comments.

Method

This preliminary study deploys thematic analysis to examine how members of the r/DataHoarder subreddit conceptualize, describe, and enact their data hoarding practices. Thematic analysis allows researchers to qualitatively surface ideas generated by individuals, communities, or institutions. As noted in Braun and Clarke (2006), theme identification does not necessarily correlate with a quantitative measure such as prevalence in the dataset. We follow Braun and Clark's six phases of thematic analysis: familiarization with data; generating initial codes; searching for themes; reviewing themes; defining themes; and reporting findings (Braun & Clarke, 2006, p. 87-93). We employ thematic analysis alongside theoretical commitments from CoPs research, whose concepts serve as guide points for framing themes deductively, while still offering space for inductive themes to emerge.

Reddit has been studied using thematic analysis within scholarly fields ranging from gender and sexuality studies to addiction medicine (Proferes et al., 2021; Graves et al., 2022; Graham & Rodriguez, 2021). When applying thematic analysis towards Reddit as a site with multiple CoPs, we note past work that addresses the complex, often socially contentious practices this platform affords users, especially as a means to examine topics which they might face criticism or institutionalized intervention (Lundeen et al., 2024; Maxwell et al., 2020).

Data collection

We began data collection by gathering the subreddit's most popular posts throughout its history (as of August 2024). A cutoff point of 2000 votes (tallying upvotes and downvotes) was chosen for this initial analysis, resulting in 170 total posts included in the analysis. Posts include a range of

text-based original posts, cross-posting of links to outside news articles, photographs, and memes. The links to these 170 posts were documented within a shared spreadsheet for purposes of analysis. Post titles, authorship (reddit username) and select excerpted comments were also collected for select posts but will not be shared through direct quotes or reproduced in the analysis here. While this information is available on the public web, we follow guidance from Proferes et al. (2021) to respect the privacy of social media users, especially as this initial analysis is not yet able to address the extent of sensitive and personal information shared within these posts or the potential for users' identities to be triangulated through other posts and comments made on reddit.

Data analysis

Following the iterative process described in Braun and Clarke (2006), the researchers deployed constant comparative coding to identify both emergent themes, while also working to create shared definitions for the themes which emerged. Work began with coding the top ten percent of the total posts within the dataset (n=17 posts), ranked by vote-count. For each post, we independently analysed the text, images, comments, and interactions to develop an initial set of codes that would inform the development of overarching themes. Upon identifying these codes, we then developed an initial set of themes that applied to these first 17 posts, before proceeding to validate these themes against the subsequent top posts to ensure that no other salient codes emerged. As an intermediate step in this iterative process, a set of high-level codes emerged. Sample codes included: materiality; value; and preventing loss. We then reviewed the remaining posts, each focusing primarily on identifying posts and comments aligned with one of these themes. A final round of discussion and refinement led to the identification of two key, overarching themes which we characterize and discuss in greater detail below.

Findings

Based on this initial analysis, we identify a set of key themes that define the r/DataHoarder vision of digital curation as a practice.

Theme (1) materiality of hoarding

Many of the top posts in r/DataHoarder highlight the materiality of hoarding, through imagery of storage media (e.g., photos of hard drives) as well as discussion of services and costs (e.g., fees for cloud storage or high-bandwidth ISP usage). Within this space, talking about and demonstrating the capacity for data storage becomes a central signifier of one's membership in the community. Practices enacted around materiality include sharing information on acquiring storage (notices of hard drives on sale), as well as sharing advice about in-home storage set-ups. We also take materiality to include the processes of acquiring data or digital objects, such as offline downloads of '*all of Wikipedia*,' and comments often signaling the creation of such a collection as a telltale sign of digital hoarding. Materiality as presented by r/DataHoarder posts often conflates data and digital objects, and ultimately emphasizes having control over digital materials through such storage.

Subtheme 1.1 fetishization and bragging

An additional facet of this focus on materiality emerges in the fetishization of cutting-edge hardware and technologies such as multiple petabyte hard drives. One post garnering significant comments and discussion includes a photo of the hard drives storing the data used to generate the first ever photographic image of a blackhole in 2019. Indicative of the fixation not on the black hole photograph itself, but instead of the storage devices, commenters metaphorically fetishize the hardware via suggestions of sexual pleasure in response to seeing these images. Another popular post presents the message received from a cloud hosting platform, that the user has exceeded their storage capacity, representative of a trend in which members brag about surpassing limitations on data plans. Not only are members concerned with storage capabilities, but the

aspect bragging about capacities of hardware or cloud-based storage, often fetishizing the ability to store previously unimaginable amounts of information.

Subtheme 1.2 nostalgia

In contrast to the fetishization of ever-newer, bigger storage, we observe how nostalgic leanings also shape the view of digital materials. One post describes a '*collapse-proof laptop*,' with comments discussing the utility of smaller, often obsolete formats and storage hardware as a mode to counteract anticipated corporate and institutional failures harkening back to a '*simpler time*' of computing's past. This aligns with a survivalist mentality of '*internet collapse preppers*', (r/DataHoarder, 2017) and the need to create tools which function without pervasive online connections and cloud services (Grandhi et al., 2020). In the same post, members suggest retaining copies of video games such as Tetris and Quake II which reflect low data demand yet represent high yield entertainment. The role of software from one's youth and a tendency to focus on collections around retro-gaming is also indicative of this nostalgic turn, as seen in another post referencing collecting rare data tapes from Nintendo 64's development. The interest in materiality therefore often overlaps with this nostalgia for obsolete hardware and formats which recalled simpler ecosystems for hoarding data and digital objects.

Theme (2) call-to-action

The call-to-action is one of the themes we largely anticipated at the outset of this study, as it aligns with r/DataHoarder activities related to the January 6th United States Capitol. Since the subreddit's description focuses largely on its role as a space for '*like-minded individuals to exchange strategies, war stories, and cautionary tales of failures*,' (r/DataHoarder, 2017) these moments that we identify as calls-to-action are rather exceptional in relation to the day-to-day discussions on the subreddit. Importantly, these interventions provide the rare opportunity for data-hoarders to directly collaboration and take collective action. Many calls-to-action revolve around urgent collecting of materials directly at risk due to corporate decisions for removal, as well as political turmoil or threats of censorship. For instance, an additional example from our dataset calls on the r/DataHoarder community to help 'backup' the social media and website of Hong Kong Stand News after their headquarters were raided and staff arrested by national security police in 2021.

Subtheme 2.1 vigilance

Related to direct calls-to-action, we observe that participants in r/DataHoarder express the need to remain vigilant, or '*stand watch*' in the face of potential future threats. This includes activities such as: sharing details on proposed changes to legislation affecting data transfer, privacy, and encryption; news of legal cases won or lost by perceived adversaries (namely large companies like Apple, Microsoft, YouTube, Yahoo, as well as Reddit itself), as well as ongoing legal issues experienced by the perceived ally, non-profit Internet Archive.

Subtheme 2.2 vigilantism

Aligned with the overarching theme of call-to-action, the motivations driving many individuals to act often appear related to an ideal of vigilante justice. This surfaces in direct ways, as r/DataHoarder posters position their work as countering the aims of unethical and unjust companies. The r/DataHoarder community therefore justifies the use of hacking, piracy, and extralegal means for taking data out of the hands of these corporations, and hoarding is perceived as contributing to the public good of making information free, open, and accessible. Several examples relate to specific projects in open science and open culture, including one post linking a repository of thousands of scientific articles on COVID-19, made openly available through torrents and decentralized web links, referencing deceased hacktivist Aaron Swartz. Beyond these direct references to piracy and hacktivism, the vigilante ethos arises across the subreddit's discussions, for instance a post offering a 'bounty' for recovery of a digitized episode of a prominent 1980s TV talk show.

Discussion

The analysis of these themes reveals where the activities and capabilities of data hoarding converge with, or diverge from, digital curation within memory institutions. Compared to the DCC lifecycle, the practice of data hoarding begins with a focus on storage, rather than pre-ingest activities of appraisal and selection. Data-hoarders believe everything that you can get your hands on can and should be hoarded, and the only limit is how many TB hard drives you can buy. An interesting point of comparison is also seen in how hoarding addresses the DCC lifecycle's activities of '*create or receive*'; for data hoarding practice, these activities are much more active than passive, requiring considerable workarounds as data-hoarders seek out and capture materials through illicit means. Data-hoarders take a unique perspective on '*preservation actions*' compared to institutional practices concerned with migration, emulation, and fixity checks. Instead, data-hoarders preserve through torrents, seeding, and re-seeding, embracing the idea of, *'lots of copies keep stuff safe*'. From this perspective, access and preservation are much more closely aligned, or even seen as the same activity, enabled by their work outside of organizational policies and concerns with copyright.

We also observe that data hoarding practices are often explicitly driven by personal motivations of nostalgia and political stances. This presents an interesting contrast to curation practice in organizational settings, which can also, equally be influenced by personal interests of curators, though this may not be openly acknowledged due to a perceived need for institutional neutrality. The way data hoarding outwardly accepts the personal nature of memory work and collecting is a welcome change.

The key themes described above represent an initial analysis of how the r/DataHoarder community defines and constructs practices of data hoarding. We note that this preliminary work has several limitations, particularly in the selection of a dataset based on top posts. Many of the more everyday routines, as well as processes of enrollment and enculturation may not be present in our selected dataset. In future work we aim to study the *'long tail*' dataset of posts with significantly less engagement and employ ethnographic methods to generate rich descriptive observations to understand the values and ethics driving this practice. Additional posts studied being to reveal processes of identity formation through r/data-hoarder, seen in posts with sentiments such as *'this is why I hoard.*' This identity can also become all-consuming and challenging for posters who subsequently describe a need to stop hoarding or describe how hoarding affects the relationship with their significant other. The role of gender within this community, and the prevalence of pornography as a prevalent object of hoarding is another topic, we have not been able to analyze in detail in this preliminary work. Through subsequent phases of this study, we also hope to address if or how data hoarding distinguishes between data and other digital objects (like the articles, books, media, and games that are often the subject of hoarding). Ultimately, by comparing the work of data-hoarders with the work of curation in memory institutions, and what each is able to target in their collections, we aim to understand what gaps and absences remain.

Conclusion

This work-in-progress highlights how r/DataHoarder offers a view of digital curation practice focused on the materiality of digital storage, and calls-to-action that raise public awareness of the potentially irrevocable loss of digital information. This presents useful counterpoints to commonly referenced and standardized practices of memory institutions by focusing on how communities conceptualize, practice, and develop their own ethical framework toward sustained access of digital materials. In future work we hope to address many questions and topics out-of-scope for this paper, such as expanding to other communities, comparing across communities, and applying both etic and emic coding to community posts.

Acknowledgements
The authors thank the anonymous reviewers for their comments and suggestions.

About the authors
Emily Maemura is Assistant Professor at the School of Information Sciences, University of Illinois Urbana-Champaign, USA. She completed her PhD at the University of Toronto's Faculty of Information. Her research focuses on data practices and the activities of curation, description, characterization, and re-use of archived web data. She can be contacted at emaemura@illinois.edu

Travis L. Wagner is Assistant Professor at the School of Information Sciences, University of Illinois Urbana-Champaign, USA. They completed their PhD at the University of South Carolina's School of Information Science. Their research focuses on the sociotechnical relationships between archival artifacts, digital curation technologies, and community representation, with an emphasis on its impact on LGBTQIA+ communities. They can be contacted at wagnert@illinois.edu

References
Basu, T. (2021, January 8). The scramble to archive Capitol insurrection footage before it disappears. MIT Technology Review. https://web.archive.org/web/20210108224223/https://www.technologyreview.com/2021/01/08/1015929/archive-capitol-insurrection-trump-maga-footage/

Bowker, G. C., & Star, S. L. (2000). Sorting Things Out: Classification and Its Consequences. MIT Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Britt, B. C., Britt, R. K., Hayes, J. L., & Oh, J. (2022). Continuing a community of practice beyond the death of its domain: Examining the Tales of Link subreddit. Behaviour & Information Technology, 41(1), 159–180. https://doi.org/10.1080/0144929X.2020.1797173

Brown, J. S., & Duguid, P. (1991). Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation. Organization Science, 2(1), 40–57.

Brown, J. S., & Duguid, P. (2001). Knowledge and Organization: A Social-Practice Perspective. Organization Science, 12(2), 198–213. https://doi.org/10.1287/orsc.12.2.198.10116

Chapman, E. M. (2023). ARCHIVING THE INSURRECTION: THE CASE OF R/DATAHOARDER. AoIR Selected Papers of Internet Research. https://doi.org/10.5210/spir.v2022i0.12987

Cox, A. (2005). What are communities of practice? A comparative review of four seminal works. Journal of Information Science, 31(6), 527–540. https://doi.org/10.1177/0165551505057016

Dallas, C. (2016). Digital curation beyond the "wild frontier": A pragmatic approach. Archival Science, 16(4), 421 - 457. https://doi.org/10.1007/s10502-015-9252-6

Grandhi, S. A., Plotnick, L., & Hiltz, S. R. (2020). An Internet-less World? Expected Impacts of a Complete Internet Outage with Implications for Preparedness and Design. Proc. ACM Hum.-Comput. Interact., 4(GROUP), 03:1-03:24. https://doi.org/10.1145/3375183

Graves, R. L., Perrone, J., Al-Garadi, M. A., Yang, Y.-C., Love, J., O'Connor, K., Gonzalez-Hernandez, G., & Sarker, A. (2022). Thematic Analysis of Reddit Content About Buprenorphine-

naloxone Using Manual Annotation and Natural Language Processing Techniques. Journal of Addiction Medicine, 16(4), 454–460. https://doi.org/10.1097/ADM.0000000000000940

Haythornthwaite, C., Kumar, P., Gruzd, A., Gilbert, S., Esteve del Valle, M., & Paulin, D. (2018). Learning in the wild: Coding for learning and practice on Reddit. Learning, Media, and Technology, 43(3), 219–235. https://doi.org/10.1080/17439884.2018.1498356

Hills, M. (2015). The expertise of digital fandom as a 'community of practice': Exploring the narrative universe of Doctor Who. Convergence, 21(3), 360–374. https://doi.org/10.1177/1354856515579844

Hudgins, W., Lynch, M., Schmal, A., Sikka, H., Swenson, M., & Joyner, D. A. (2020). Informal Learning Communities: The Other Massive Open Online 'C.' Proceedings of the Seventh ACM Conference on Learning @ Scale, 91–101. https://doi.org/10.1145/3386527.3405926

Kwon, K. H., Yu, W., Kilar, S., Shao, C., Broussard, K., & Lutes, T. (2020). Knowledge Sharing Network in a Community of Illicit Practice: A Cybermarket Subreddit Case. Proceedings of the 53rd Hawaii International Conference on Systems Sciences, 2731–2740. http://hdl.handle.net/10125/64076

Lave, J., & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511815355

LeBlanc, Z., Janco, A., Wermer-Colan, A., Dombrowski, Q., Kijas, A., Majstorovic, S., Strong, D., & Peaslee, E. (2022). A Conversation with the Organizers of Saving Ukrainian Cultural Heritage Online (SUCHO). Journal of Library Outreach and Engagement, 2(1), Article 1. https://doi.org/10.21900/j.jloe.v2i1.969

Lundeen, L. A., McCall, J. R., Bradshaw, A. S., LeBlanc, E. L., & Humphrey, L. (2024). Digital Catharsis or Harmful Exposure? A Thematic Analysis of Self-Directed Violence Reddit Posts. Social Media + Society, 10(3), 20563051241263562. https://doi.org/10.1177/20563051241263562

Maxwell, D., Robinson, S. R., Williams, J. R., & Keaton, C. (2020). "A Short Story of a Lonely Guy": A Qualitative Thematic Analysis of Involuntary Celibacy Using Reddit. Sexuality & Culture, 24(6), 1852–1874. https://doi.org/10.1007/s12119-020-09724-6

McMeekin, S., & Currie, A. (2022). DPC Competency Audit Toolkit Guide (1st ed.). Digital Preservation Coalition. https://doi.org/10.7207/dpccat22-01

Moran, J., Feltham, F., & Love, V. (2019). Building an Aotearoa New Zealand-wide Digital Curation Community of Practice. International Journal of Digital Curation, 14(1), Article 1. https://doi.org/10.2218/ijdc.v14i1.638

National Digital Stewardship Alliance (NDSA). (2019). 2019 Levels of Digital Preservation Matrix. https://osf.io/2mkwx/

Ogden, J. (2022). "Everything on the internet can be saved": Archive Team, Tumblr, and the cultural significance of web archiving. Internet Histories, 6(1–2), 1–20. https://doi.org/10.1080/24701475.2021.1985835

Panek, E. T. (2022). Understanding Reddit. Routledge.

Peltzman, S., Dietz, B., Butler, D., Walker, P., Farrell, J., Arroyo-Ramirez, E., Macquarie, C., Bolding, K., Helms, A., Venlet, J., Cobourn, A., Watson, C., Taylor, A., & Henke, K. (2022). Levels of Born-Digital Access. https://doi.org/10.17605/OSF.IO/R5F78

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. Social Media + Society, 7(2), 20563051211019004. https://doi.org/10.1177/20563051211019004

r/DataHoarder. (2017, July 30). A quick Datahoarder FAQ [Reddit Post]. https://web.archive.org/web/20171211044900/www.reddit.com/r/DataHoarder/comments/6qf716/a_quick_datahoarder_faq/

Rios, F., Lassere, M., Ruggill, J. E., & McAllister, K. S. (2020). Sustaining Software Preservation Efforts Through Use and Communities of Practice. International Journal of Digital Curation, 15(1), Article 1. https://doi.org/10.2218/ijdc.v15i1.696

Walker, D., Nost, E., Lemelin, A., Lave, R., & Dillon, L. (2018). Practicing environmental data justice: From DataRescue to Data Together. Geo: Geography and Environment, 5(2), e00061. https://doi.org/10.1002/geo2.61

Wenger, E. (1998). Communities of Practice: Learning as a Social System. Systems Thinker, 9(5), 2–3.

Wenger, E. (2000). Communities of Practice and Social Learning Systems. Organization, 7(2), 225–246. https://doi.org/10.1177/135050840072002

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Graphicon evolution on Bilibili: a historically informed socio-technical approach

Yiqiong Zhang, Susan C. Herring, Yanmin Wu, and Rongle Tan

DOI: https://doi.org/10.47989/ir30iConf47194

Abstract

Introduction. Digital technologies have transformed the way we communicate, yet research on digital communication has often focused on either social or technical factors in isolation and has generally lacked a historical perspective.

Method. This study integrates a historical perspective with a socio-technical approach to examine the trends and factors driving the evolution of graphicons (kaomojis, emojis, and stickers) on the Chinese social media platform Bilibili.

Analysis. The analysis involves two parts: first, identifying graphicon evolution patterns from a 14-year corpus and irregularities in relation to a predicted rise-and-fall model, and then using digital archaeology to reconstruct the history and usage of graphicons on Bilibili, classifying them into social or technical categories.

Results. Findings show that only kaomoji usage follows a clear rise-and-fall trend, while trends for emoji and sticker use are less clear. The evolution of graphicons is influenced by both social factors, like user discourse behaviour and graphic arts, and technical factors, such as design and input interfaces.

Conclusion. The historically informed socio-technical approach offers a feasible framework for future studies of digital phenomena. Its quantitative aspect allows the identification of longitudinal patterns and irregularities in the data, while its qualitative aspect enables greater depth through considerations of the events underlying the data.

Introduction

Digital technologies have dramatically transformed communication, enabling the creation, sharing, and interpretation of visual content in more dynamic and multimodal ways across diverse platforms and at unprecedented speeds. As part of this transformation, graphicons – graphical icons used in text-based computer-mediated communication (Herring & Dainas, 2017) – have emerged as particularly powerful tools, significantly reshaping how people express emotions, share ideas, and convey cultural nuances in digital spaces. It is thus crucial to understand how graphicons evolve and the implications of that evolution for technology-mediated communication.

Graphicons, as technical artefacts used in computer-mediated communication (CMC), comprise both social and technical dimensions. Studies thus far, however, have focused mainly on social perspectives, such as graphicon meanings and functions (e.g., Herring & Ge-Stadnyk, 2024; Logi & Zappavigna, 2021), gender and age usage patterns (Koch et al., 2022), and cultural influences (e.g., Chik & Vásquez, 2017; Zhang et al., 2021). That is, the focus has primarily been on how people use graphicons to communicate in various social contexts, while less attention has been paid to the technical factors that influence the use and evolution of graphicons. This oversight mirrors broader trends in digital communication research, where the role of technological aspects such as algorithmic design and platform interfaces has been underexplored, despite their significant impact on shaping how communication unfolds in digital environments (van Dijck, 2013).

Both perspectives are important when seeking to understand the factors that condition changes in graphicon use over time. Broadly, graphicons have evolved from simple, schematic ASCII (American Standard Code for Information Interchange) emoticons composed of keyboard symbols to more representational emojis, and further to diverse and colourful stickers on social media platforms. Konrad et al. (2020) propose that the evolutionary trends of these graphicons follow a rise-and-fall pattern, where a newer category of graphicon gradually replaces its predecessor. However, their proposal was inferred from synchronous, self-reported data. It remains largely unexplored how well their model aligns with actual graphicon usage and what mechanisms drive the progression in practical contexts of use.

The present study seeks first to identify trends in graphicon evolution on the Chinese social media platform Bilibili and then to discover the factors that contribute to those trends. It adopts a historically informed socio-technical approach, incorporating methods from digital archaeology – the use digital technology to investigate the past – to reconstruct the history and usage of graphicons on Bilibili. Through analysing a 14-year longitudinal corpus of Bilibili comments, we identify several key social and technical factors that arguably contributed to the observable trends and irregularities. More generally, this approach provides significant explanatory power for understanding graphicon evolution and can be applied to the study of digital communication across other platforms and cultural contexts.

Literature review

Graphicon evolution

Graphicon studies have focused mainly on the meanings and functions of individual categories of graphicons (typically emoticons, emojis, or stickers) (see the review in Tang & Hew, 2019) rather than considering graphicons in relation to one another as part of a larger ecology, as recommended by Herring and Dainas (2017). Even fewer studies have explored the changes that take place over time within this ecology.

An exception is Konrad et al. (2020), who posited that graphicons typically pass through three phases: an early phase, a peak phase, and a phase of decline or conventionalisation. In Western contexts, Konrad et al. (2020) situated emoticons in the third phase, based on their reduced frequency and their conventionalised use as punctuation; emojis in the second, peak phase; and

stickers in the first, earliest phase. They also predicted that the trajectory would be more advanced in Asian countries, with stickers in ascendance and emojis starting to decline. Their predictions were based on interviews and surveys conducted with English-speaking Facebook messenger users rather than longitudinal data, however.

Studies with actual longitudinal data so far have confirmed that emoji have overtaken emoticons in the West (Pavalanathan & Eisenstein, 2016). As emoticons have declined in frequency of use, their functions have become conventionalised as a type of punctuation (Markman & Oshima, 2007; Provine et al., 2007), with the conventionalisation further demotivating their usage (Albert, 2020; Konrad et al., 2020). A longitudinal study in a Chinese context that examined the frequencies of kaomojis (Japanese-style emoticons), emojis, and stickers found only partial support for the prediction of Konrad et al. (2020), however. Kaomojis exhibited a rise-and-fall pattern, but their decline was not accompanied by an increase in emojis or stickers (Zhang et al., 2022).

The discrepancy between model-based predictions and actual data highlights the need for further investigation into the factors influencing these trends. Konrad et al. (2020) propose that popularity (frequency of use) and shift in pragmatic function are the primary drivers of the transition between graphicon categories. However, these could be seen as effects rather than as causes of the rise and fall of graphicon categories. It remains to be determined what factors, especially technical ones, drive their evolution.

Additionally, although an increasing number of new graphicon designs within each graphicon category are available, little is known about how each unique graphicon is used in practice. Several quantitative studies have identified the top 20 most frequently used emojis in different corpora (e.g., Liu et al., 2020; Wang et al., 2024), but little is known about the use of less frequently occurring emojis. Zhang et al. (2022) found in their longitudinal corpus that the trends in graphicon types differ significantly from those for graphicon tokens; however, no reasons for this discrepancy were given.

Graphicons on Chinese social media and Bilibili

Graphicons on Chinese social media constitute a unique ecosystem with distinctive designs and usage. The graphicons incorporate the logographic features of Chinese characters (Li & Zhu, 2019; Zhang et al., 2022) and convey rich cultural messages from Chinese idioms and traditions (de Seta, 2018; Zhang et al., 2021). The culturally enriched designs of graphicons on Chinese social media are preferred over Unicode emojis (e.g., 😀) by users of different platforms (de Seta, 2018). Chinese users tend to use emojis more frequently in contexts involving family and social relationships compared to their Western counterparts, who also use emojis extensively in professional and casual settings (Guntuku et al., 2019). Although several studies have pointed out that stickers are now very popular among Chinese social media users (e.g., de Seta, 2018; Ge, 2020; Ge-Stadnyk et al., 2025; Sandel et al., 2019), we have been unable to find comparable information on the frequency of sticker use.

Graphicons on Bilibili are distinctive in several ways that make the platform a valuable case study for graphicon usage. Bilibili is one of the most popular social media platforms in China, featuring high stickiness and interactivity (Gao, 2021; Sun et al., 2023; Tian, 2021). As we argue here, graphicons play a significant role in user interaction, cultural expression, and community building. Bilibili has released platform-specific sets of emojis and stickers that have been overwhelmingly popular – for instance, a TV set icon that is the symbol of Bilibili, and the Popular Words Series, which are graphic representations of selected popular verbal expressions from comments or '*danmu*' (short messages overlaid on the video itself) (Zhang et al., 2022). These platform-specific graphicons, which are associated with Bilibili's fandom cultures, enhance user engagement through expressive visual communication, allowing users to indicate their membership and build a sense of identity within the platform community (Wu et al., 2018). The platform also monetises

graphicons by offering exclusive packs for purchase, encouraging users to invest more in their online presence (Yang, 2021).

A socio-technical perspective on graphicon evolution

The development and widespread adoption of graphicons, as with other social media affordances, are the result of ongoing interactions between the digital platform's capabilities and the social needs and practices of its users. This interplay is what Trist (1978) called socio-technical systems theory, which emphasises the interaction of social and technical subsystems. The social subsystem focuses on human factors such as people, relationships, and norms, while the technical subsystem involves tools and technologies, such as hardware and software, required for tasks (Bostrom & Heinen, 1977; Kling, 2007). Both aspects are crucial for understanding digital communication.

A socio-technical perspective has been found useful in exploring the evolution of digital communication, which is marked by the mutual influence of social practices and emerging technologies. For instance, it has shown how digital maps transitioned from traditional cartographic methods to web-native media forms that reshape social practices and spatial awareness (Hanchard, 2024). Studies in this line demonstrate the potential of the socio-technical approach to shed light on the complexity in evolving digital communication.

To the best of our knowledge, this perspective has not been explicitly applied in graphicon studies, although the focus of most existing graphicon research could be categorised as either social or technical. Studies from the social perspective explore topics such as the meanings and functions of graphicons (Herring & Ge-Stadnyk, 2024; Logi & Zappavigna, 2021), usage patterns in relation to user demographics and profiles (Koch et al., 2022), and cultural influences on graphicon use and interpretation (e.g., Chik & Vásquez, 2017; Zhang et al., 2021). In contrast, the technical perspective has been considerably less studied. Yet technical factors such as the design and availability of graphicons clearly have an impact on their usage. For instance, Unicode, as a technical standard, distinguishes emojis from emoticons and other iconographic formats, enabling their use across devices, operating systems, and Internet platforms, thereby facilitating their widespread circulation (Berard, 2018). The layout of the emoji keyboard has an impact on the time required for emoji selection (Pohl et al., 2017), thereby affecting its usage. Stickers, unlike emojis, are not supported by Unicode or standardised formats, which could hinder the development of widespread conventions of sticker use (Tagg & Lyons, 2021). While social and technical perspectives both offer valuable insights into graphicon usage, studies investigating their interactions within the same longitudinal corpus are notably lacking.

Research questions

This study extends the socio-technical perspective by integrating a historical analysis of both social and technical influences in a large corpus of graphicons. Specifically, we explore the following research questions: 1) What are the trends over time of the relative frequencies of Chinese graphicons on Bilibili? 2) What irregularities in the trends are observed when compared to the rise-and-fall model of graphicon evolution proposed by Konrad et al. (2020)? and 3) What socio-technical factors in the platform's history explain the trends in the Bilibili corpus, particularly the irregular frequencies that deviate from the predictions of Konrad et al. (2020)?

Methodology

Data collection

Our corpus consists of 14 years of longitudinal comment data collected from the Bilibili platform. Bilibili is a video-sharing platform that, like YouTube, allows users to post comments below the videos and features short danmu messages that are overlaid on the video itself.

The Bilibili platform was chosen for several reasons. Firstly, it is one of the most popular Chinese social media platforms, with average monthly active users reaching 0.315 billion (around one-fifth

of the Chinese population) in the first quarter of 2023 (https://www.Bilibili.com/read/cv24070769/, retrieved July 9, 2024). Secondly, the comments can include kaomojis, emojis, and stickers. Lastly and most importantly, Bilibili preserves a historical record of the comments posted below the videos, including the graphicons in the comments, and the comments can be captured automatically.

The data consist of comments and replies to comments from the channel of Bilibili's annual Spring Festival Gala Show (hereafter, the Bilibili show) (https://space.Bilibili.com/1868902080, retrieved April 5, 2022). This channel was chosen because it is the only one that includes consecutive comments dating back to 2010. The show consists of a mash-up video of content provided by professional users to celebrate the Chinese New Year, and it is released on the eve of the Chinese New Year. Besides the show videos, the Bilibili show channel includes videos related to the gala show, such as trailers, teasers, and outtakes. Like the Bilibili show videos, these videos include older comments and replies and thus are included in our data.

Comments and replies from all 54 videos available in the channel, covering the years from 2010 to 2023, were captured and stored in March 2023 using Python and the Scrapy tool. A total of 992,856 messages (including both comments and replies) were collected. Danmu messages were excluded from the data due to their limited use of graphicons.

Data analysis

The analysis was conducted in two stages. To obtain the trends of kaomojis, emojis, and stickers in use, we first identified their occurrence in the messages using a semi-supervised process of deep learning (Qin et al., 2019) and manual identification and calculated their frequency. The frequencies were normalised based on the total number of messages each year. We also applied Voyant (https://voyant-tools.org/), a web-based text analysis tool for analysing word frequencies and trends in comments. We derived patterns for the corpus as a whole, as well as identifying deviations from the model predicted by Konrad et al. (2020).

In the second stage, to account for the patterns in the frequencies, we adopted a historically informed socio-technical approach. We conducted digital archaeology to reconstruct the history and usage of graphicons on Bilibili, based on which we identified factors that could plausibly be related to irregularities and classified them into social or technical categories. Digital archaeology refers to 'the use of digital technology to investigate and communicate the past' (Morgan, 2022, p. 214); in this case, the past being examined is also digital. Specifically, our search for historical information included the following major sources: 1) information available on Bilibili for the timeline of events, platform practices, and culture; 2) annual Bilibili financial reports for user numbers and profiles; 3) media reports about Bilibili for major events in its development and its social impact; 4) the internet archive Wayback Machine for old Bilibili webpages; and 5) questions and answers about Bilibili on platforms such as Zhihu (Chinese Reddit) for graphicon input settings and graphicon accessibility.

We recognise that our backgrounds as researchers influence our analysis of Bilibili's user communities and platform dynamics. Our approach is informed by cross-disciplinary perspectives from computer-mediated communication, linguistics, and information science. Further, the third and fourth authors are native Chinese speakers who are users of Bilibili, and the first and second authors have previously researched Bilibili and other Chinese social media platforms. While we believe these perspectives enrich our analysis, we acknowledge that our interpretations may be shaped by our familiarity with Chinese cultural contexts. Concerning data collection and use, the study followed the platform's guidelines for data access, scraping, and usage for academic research.

Findings

This section presents the frequencies of graphicon usage on Bilibili over time and explores how irregularities in the frequencies can be explained with reference to key historical events associated with the platform, which we classify into social or technical categories.

Trends in graphicon evolution on Bilibili

The normalised frequencies of all graphicon tokens for each of the 14 years are displayed in Figure 1.

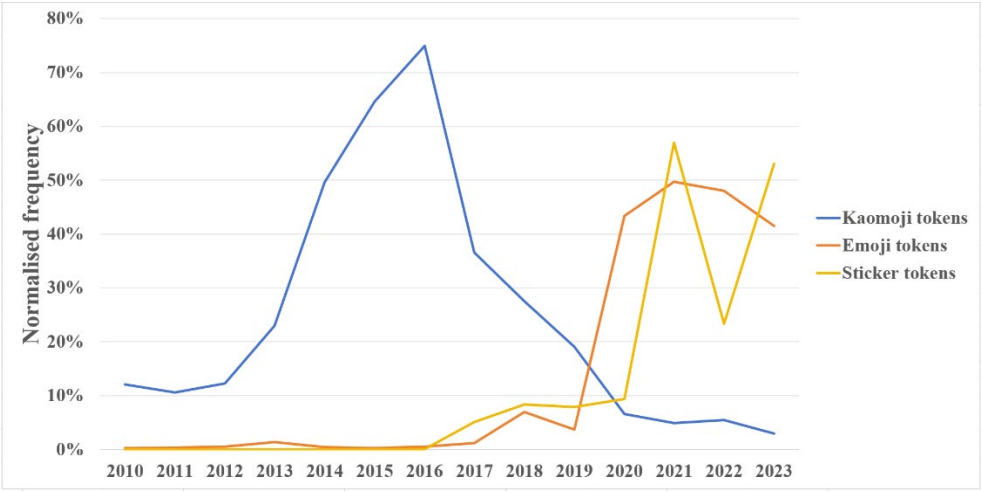


Figure 1. Frequencies of graphicon tokens by year (normalized in relation to the number of messages for each year)

These trends provide partial support for the evolutionary trajectory of graphicons proposed by Konrad et al. (2020), particularly as regards kaomoji. Deviations from Konrad et al.'s (2020) predicted rise and fall of each graphicon (emoji replacing emoticons (kaomojis) and stickers replacing emojis) include the following:

1) Although kaomoji rose and fell from 2010 to 2019, the decrease in kaomoji was not replaced by emojis or stickers, leading to a significant drop in graphicon use in 2017.
2) Emojis were not popular on Bilibili between 2010 and 2019, despite being popular on other social media platforms in China during that same period. Emoji tokens increased significantly in 2020 but remained level for the following three years.
3) Stickers, which first appeared on Bilibili in 2016, showed a steady increase except for a dramatic spike in 2021.

History of the platform

As part of the findings from our digital archaeology, a brief overview of the platform's history is essential to understand the socio-technical factors contributing to the irregularities in graphicon evolution. Bilibili was founded in 2009 as a non-profit video-sharing community for Anime, Comics, and Games (ACG) fans (Sun et al., 2023). The platform derives its name from the popular anime *A Certain Magical Index*, where the protagonist Mikoto Misaka's electric attack produces a sound humorously referred to as *Bilibili*, which later became the nickname given to her by the male protagonist (https://www.Bilibili.com/read/cv154805/, retrieved June 4, 2024). In 2013, Bilibili received its first investment from Tencent, was listed on the NASDAQ stock index in 2018, and was listed on the Hong Kong Stock Exchange in 2021. Along with its increasing commercialisation, Bilibili has seen changes in its policies, content, and the number and profiles of its users.

Bilibili released open registration in 2013 with a strict membership test to familiarise the newcomers with the platform's culture and practices. Only individuals who passed the test were

granted permission to interact with comments or '*danmu*'. The test was discontinued in 2015 to accommodate a broader group of users. Not long after, a paid membership was implemented in 2017, suggesting the platform's full commitment to commercialisation. Together with these developments, the video content on the platform expanded from pure ACG to pan-ACG (including general games and animations, then movies and TV programs), then further to a broad spectrum of videos. Alongside the expansion of video content categories, the main user groups also shifted from ACG fans to pan-ACG fans, and then to pan-entertainment users. By 2023, the monthly active users had increased to over 300 million. The expansion of the platform from a closed community for ACG lovers to a massive pan-entertainment site is illustrated in Table 1, alongside trends in graphicon use.

Stage	1	2	3	4
Time	2010-2012	2013-2016	2017-2019	2020-2023
Platform commercialisation	Non-profitable community for ACG	Investment from Tencent	Listed on NASDAQ in 2018	Opened to all brand partners in 2020; listed on the Hong Kong Stock Exchange in 2021
Platform policies	Time-limited registration system	Strict membership test started in 2013, but removed in 2015	Paid membership in 2017	Hardcore membership test in 2022
Video content on the platform	ACG	Moving from ACG to pan-ACG, including general games and animations, then to movies and TV programs	Non-ACG videos outnumbered ACG videos; the largest online community for professional user-generated video, vlogs, and classroom sessions	A broad spectrum of videos
User profiles	ACG fans	Pan-ACG fans	Pan-ACG fans, Generation Z	Pan-entertainment users
Graphicon trends	Kaomoji relatively stable	Kaomoji increased	Kaomojis decreased, not replaced by emojis or stickers	Platform-specific emojis and stickers increased sharply

Table 1. Bilibili history

Socio-technical factors influencing trends in graphicon use

Social factors: user discourse behaviour

User discourse behaviour refers to the patterns and tendencies in how users communicate and interact within a particular context, platform, or medium (Herring, 2004). Variations in user behaviours can contribute to the observed irregularities in graphicon usage. Specifically, message valence (positive or negative) contributed to a significant drop in graphicon use in 2017 on Bilibili.

The valence of messages was examined because the emotional tone of messages has been found to be related to graphicon usage (Tang & Hew, 2018). Through close reading of the messages in 2017, we found that the content was predominantly serious and negative due to two major events that occurred that year. Firstly, comments were dominated by intense and serious discussions about the plot design of the 2017 show, which featured a sci-fi mystery story with three different endings. Secondly, during the 2017 gala show, the website's credit system was hacked. As a result, users at advanced experience levels 5 and 6 had their credits removed and were unable to interact

on the platform, including posting comments or danmu. In reaction to these events, two distinct focuses were observed in the 2017 comments: a concentration on the plot design of the Gala Show (Extract 1), and criticism of the hackers (Extract 2). Both sets of comments were generally serious in tone and tended to avoid the use of graphicons.

Extract 1 (2017-1-30 11:25)

我看完也是这么想的，不然男主活的解决就说不通了；活下的男主没绳索没打火机去拯救一下女主是不可能的，楼里也没男主尸骨。这样解释个人感觉比较合适。 (Translation: I had the same idea after watching it. Otherwise, the resolution of the male protagonist surviving doesn't make sense. Without a rope or a lighter, it would be impossible for the surviving male protagonist to save the female protagonist, and there were no remains of the male protagonist in the building. This explanation feels more fitting to me personally.)

In Extract 1, the commenter attempted to make sense of the plot design by ruling out various possible interpretations. No graphicons were used in the comment, and several negations were used in ruling out possibilities. Graphicons have always tended to be used in playful and positive settings (e.g., Zhang et al., 2021), but they were absent in this serious discussion. Similarly, the commenters who posted messages to the hackers in Extract 2 intended for their messages to be taken seriously, so they also avoided using graphicons, which might suggest a playful tone, as highlighted in many studies (Konrad et al., 2020; Zhang et al., 2021).

Extract 2 (2017-01-28 01:26)

致敬攻击 B 站的人：你们越这么闹，我反而越爱这个网站！(Translation: A salute to those attacking Bilibili: The more you cause trouble, the more I love this website!)

The use of negation in Extract 1 prompted us to examine the negativity of the comments more systematically. We identified occurrences of the Chinese character for negation 不 [BU] in the corpus along the timeline using Voyant for analysing word frequencies and trends in comments. The normalised frequency of the negation character, as shown in Figure 2, peaked in 2017, indicating that the degree of negativity in the comments was most prominent that year. This is the relative frequency of the term in the document, calculated by dividing the raw frequency by the total number of terms in the document and multiplying by 1 million (https://voyant-tools.org/docs/#!/guide/documentterms). The seriousness and negativity of messages in 2017 plausibly explains the drop in overall graphicon usage, particularly in kaomojis, in 2017.

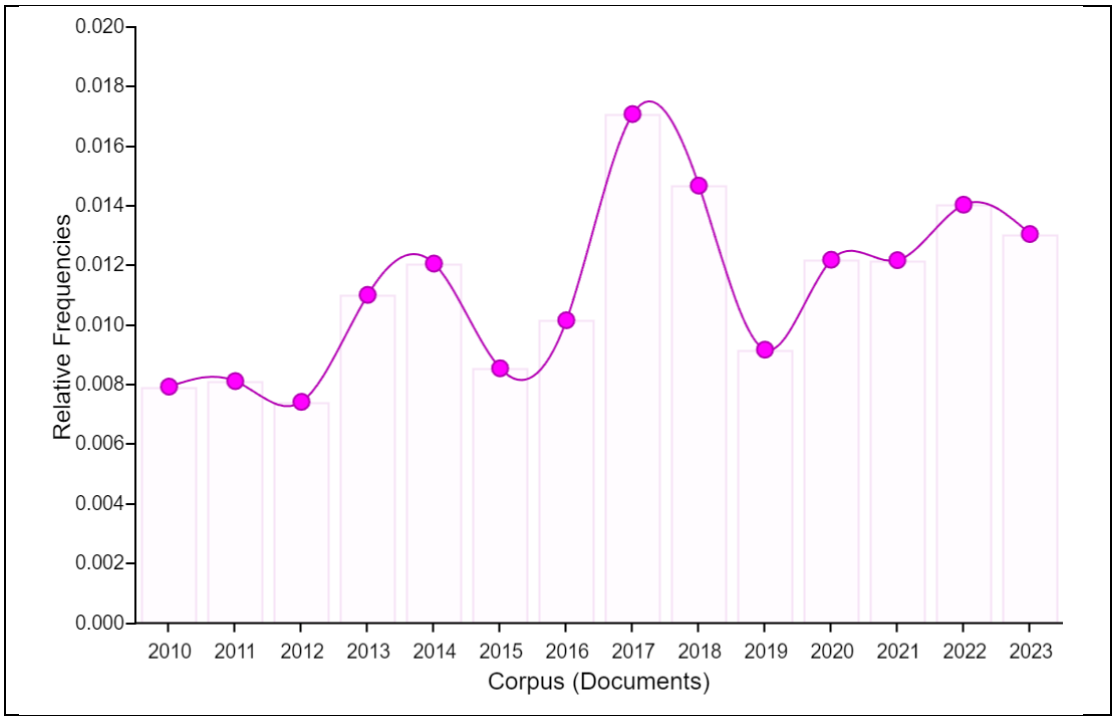


Figure 2. Frequencies of occurrence of 不 [BU] (normalized in relation to the total number of Chinese words)

A cooccurring factor that could also have contributed to the significant drop in kaomoji usage was the influx of new and non-ACG users in 2017 (as shown in Table 1), since the use of kaomojis is closely linked to ACG culture (Kaneyasu, 2022).

Another type of user discourse behaviour that plausibly impacted graphicon irregularities in our data is graphicon art. This involves using graphicons to create various shapes and designs, as illustrated in Figure 3. As emoticons gained in popularity in the 1990s, ASCII art emerged as a type of '*cyberplay*' (Danet, 2001) involving the use of the symbols from the ASCII set to craft images and designs in text-based messages (de Seta, 2018; Kozar, 1995). Graphicon art has evolved with the introduction of new graphicon forms, from ASCII symbols to emojis (Figure 3, left) and stickers (Figure 3, right). Designs composed of stickers were more popular than those with emojis in our corpus, a preference that may stem from stickers being larger and more colourful than emojis, making them a more appealing choice as graphic arts components.



Figure 3. Examples of graphicon arts in messages from 2021

Graphicon arts gained immense popularity on Bilibili in 2021, with numerous replies featuring copy-pasted designs. This surge increased the frequency of graphicons, leading to a sharp rise in sticker tokens in 2021. However, its popularity was short-lived, resulting in a temporary spike in sticker usage in that year.

Technical factors: graphicon availability

Beyond social factors, technical factors also significantly shaped the evolution of Bilibili graphicons. Graphicon availability stands out as a crucial factor in explaining their use. This refers to the presence and accessibility of graphicons on specific platforms, which in turn determines the range, variety, and ease of use of graphicons for users. This section focuses on two issues related to accessibility, graphicon design and the input interface.

Graphicon design influences graphicon usage, because visually appealing, easily recognizable, and contextually appropriate designs are more likely to be adopted and regularly used by users. This factor is especially important in the Chinese context, because users tend to prefer platform-specific graphicons that encode the community culture (de Seta, 2018; Zhang et al., 2022; Zhang et al., 2021).

Our corpus has very low frequencies of emojis and stickers until 2020, which is unexpected given their widespread use on other social media platforms well before 2020. The significant increase in emojis and stickers after 2020 is primarily due to the introduction of platform-specific sets. Notably, Bilibili released a set of Yellow Face emojis in 2019 (see examples in Table 2), with facial representations that differ slightly from those found on other social media platforms like WeChat or Weibo (de Seta, 2018; Zhang et al., 2022). The Yellow Faces set also includes emojis for the New Year's celebration, such as the New Year Blessing in Table 2. Regarding stickers, Bilibili released several free sets that have become overwhelmingly popular among users. These include the TV set, which is an iconic symbol of Bilibili, and the Popular Words Series (PWS), which are graphic representations of selected popular expressions from comments or '*danmu*' (see Table 2). These platform-specific graphicons encode Chinese characters and the discourse practices of the platform, making them among the most frequently used in Bilibili (Zhang et al., 2022).

Category	Yellow faces		Stickers	
Name	Doge	New Year Blessing	PWS: wonderful	TV set: smile
Graphicon				

Table 2. Examples of platform-specific sets of emojis and stickers

It should be noted that there was a noticeable lag between the release of graphicons and their usage in comments. Emojis and stickers released in March 2019 did not see widespread use until 2020, as most comments on the gala show were posted shortly after the event in January. This explains the increase in the frequency of emojis and stickers from 2020 to 2023, driven by the release of platform-specific sets. This rise is especially noteworthy in light of a study on emoji usage during the COVID-19 outbreak on Chinese social media, which found a substantial decline in both the proportion of posts containing emojis and the ratio of users using emojis during that period (Liu et al., 2022).

Another technical factor influencing the availability of graphicons is the input interface, which refers to the features and functionalities that allow users to type and select graphicons in their comments for submission to be displayed on the platform. We expected that the inclusion of

graphicons in the input interface would lead to an increase in their accessibility and further their usage.

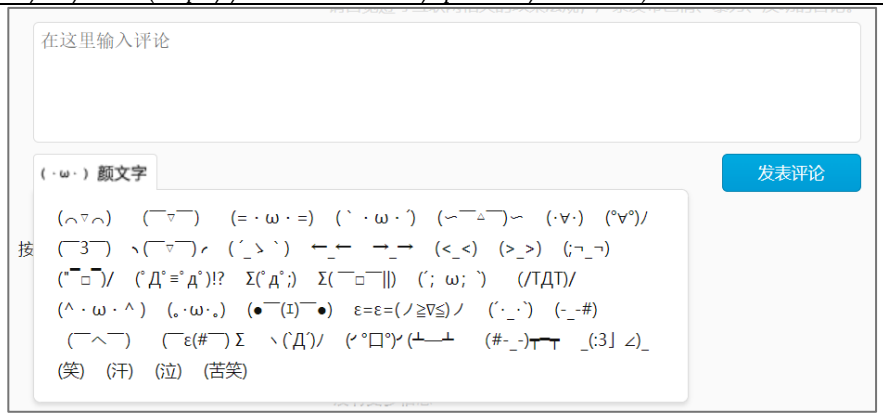
Time	Screen shot
2010-2012	No kaomoji in the input interface for comments.
2013-2014	 *From a post on Zhihu discussing the Bilibili interface design in 2014, captured 07/20/2023 (https://www.zhihu.com/question/25116069)*
2015-2018	 *From the Wayback Machine, dated 04/24/2015, captured 07/20/2023.*
2019	 *From the Wayback Machine, dated 10/02/2019, captured 07/20/2023.*
2020-2023	 *From the Wayback Machine, dated 02/03/2020, captured 07/20/2023.*

Table 3. Input interface versions on the Bilibili website from 2010 to 2023

Several successive versions of the input interface for comments on Bilibili are presented in Table 3. At the early stage of Bilibili, no graphicons were available in the input interface. The input of

kaomojis at that time required users to either manually type ASCII symbols or copy kaomojis from other sources. Kaomojis first appeared in the input interface in 2013 (see the interface of 2013-2014 in Table 3). The introduction of an auto-drop-down menu for kaomojis in 2015 (see the 2015-2018 input interface in Table 3) significantly improved input convenience, leading to a further increase in kaomoji use.

In 2019, the Yellow Faces set was introduced to the input interface, followed by an increase in emoji frequencies in 2020. The Popular Words Series and Bilibili TV series appeared in the 2020 interface, followed by an increase in sticker frequencies in 2021, 2022, and 2023. Also, in the 2020 interface, kaomojis were downplayed by using grey fonts instead of black, as in previous versions. It is also worth highlighting that the input interface did not include Unicode emojis or non-platform-specific stickers, which could be a possible reason for their usage frequencies remaining extremely low over the course of 14 years. This suggests that the (lack of) availability of a graphicon category in the input interface greatly affects its use. In support of this, we examined the top 10 kaomojis and found that they were mostly from the recommended list in the 2015-2018 interface, whereas the top emojis and stickers were from the recommended lists in the 2020 interface. It should be noted that the web interface is different from the mobile interface, which Bilibili released in February 2012. According to the mobile interface in 2023, users who made comments could either use graphicons recommended by the Bilibili interface or mobile input methods. Unfortunately, we were not able to retrieve the interface data from mobile versions.

Discussion

Based on a longitudinal corpus of comments from Bilibili, we first mapped the trends of the frequencies of kaomojis, emojis, and stickers. We then explored how these trends aligned with or diverged from the predicted rise-and-fall pattern of graphicons by Konrad et al. (2020). We investigated the history of Bilibili and uncovered both social and technical factors that plausibly influenced the evolution of graphicons.

Among the social factors, serious and negative situations in 2017 caused a marked decrease in the frequency of kaomojis in user comments (emojis and stickers had not yet become widely used on Bilibili). Previous studies using qualitative data have shown that people tend to use more emoticons in positive contexts than in negative ones (Ahn et al., 2011) and that emojis are not likely to be used in negative criticism statements (Chang, 2016). The statistics from our corpus offer the first quantitative confirmation of this tendency. While it is widely recognised that graphicons are commonly used to add a positive and playful tone to messages (e.g., Konrad et al., 2020), understanding of their meaning potential in negative contexts remains quite limited. This factor highlights the need for further exploration of graphicons in negative contexts and for developing graphicon categories to express negative meanings.

Graphicon arts represent another form of user discourse whereby users craft elaborate visual displays by blending a combination of kaomojis, emojis, and stickers, causing a notable surge in sticker usage in 2021. There has been a trend toward larger and more complex visual designs as graphicons evolved from kaomojis to emojis and then to stickers. Meanwhile, there is a growing integration of graphicons, both their design and usage, into internet meme culture. Although it is widely acknowledged that internet memes are a major source for sticker creation (e.g., Ge, 2020; Ma, 2016), our findings indicate that graphicons are also intricately intertwined with the meme culture through graphicon arts. The copy-pasted graphicon arts are frequently used in messages as memes to execute various types of speech acts such as confirm and congratulate, consistent with the concept of '*memes as speech acts*' (Grundlingh, 2018). However, graphicon arts were only popular on Bilibili in 2021, and the reasons for their decline in 2022 remain unclear. To assess the durability of the graphicon art trend, further observation with more extensive longitudinal data is necessary.

For technical factors, we proposed that the design of graphicons had a significant impact on graphicon usage on Bilibili. The most popular graphicon types were those platform-specific sets that are deeply imbued with community practices and culture, first kaomojis and then Yellow Faces and Popular Word Series. Bilibili has highly sticky user groups and community engagement through the interactivity of the platform (Gao, 2021; Sun et al., 2023; Tian, 2021). Given the sharp increase in the frequency of platform-specific graphicon sets during a period of rapid influx of new users, we propose that these graphicon sets are strategic tools employed by the Bilibili platform to boost user engagement and cultivate a sense of community belonging.

Another technical factor that had a major impact on graphicon usage on Bilibili is the availability of graphicon categories in the input interface. We noted a rise in the frequency of each category (kaomojis, emojis, and stickers) after it was featured in the comment input interface. While we were able to track the availability of the web version's input interface using archived data, we could not be certain whether the extremely low usage of Unicode emojis and non-platform-specific stickers was due to their non-appearance in the input interface. However, there is precedent for assuming that it was. In their comparative study of websites/apps for restaurant recommendations, Chik and Vásquez (2017) found that many OpenRice reviewers included emojis in their reviews, while Yelp reviewers did not, because OpenRice provided users with the option to use emojis, whereas Yelp did not.

Conclusion

To the best of our knowledge, this is the first empirical case study to explore the factors driving graphicon evolution based on actual longitudinal data. Our longitudinal analysis uncovered both social factors, such as user discourse behaviour and graphicon arts, and technical factors, including graphicon design and input interfaces, that plausibly played a role in shaping graphicon evolution on Bilibili. Findings from the case study should be validated with other topics and platforms, however, as our analysis was limited to the Bilibili channel for the Chinese New Year show.

From a methodological perspective, building on our digital archaeological investigation of the factors influencing graphicon use, we expanded the socio-technical approach into a historically informed socio-technical approach for studying graphicon evolution. This method uncovered insights and connections that might otherwise have never been made previously, emphasising the value of a contextually and historically enriched approach. Its quantitative aspect allows the identification of longitudinal patterns and irregularities in the data, while its qualitative aspect enables greater depth through consideration of the events underlying the data. In the present study, the historical aspect of this approach focused on understanding the temporal context in which graphicons have emerged, evolved, and declined, recognising their development as deeply intertwined with the broader social context and technological advancements on Bilibili. However, we are aware that the historical factors discussed in this article are not exhaustive. For instance, it is important to consider how specific hardware and mobile operating systems might influence the use of graphicons. Nonetheless, the historically informed socio-technical approach offers a feasible framework for future studies of graphicon evolution, as well as of other digital phenomena. It encourages researchers to consider historical and contextual factors for more comprehensive analyses. For platform designers and developers, our findings highlight the importance of both user behaviour and technical design in shaping digital communication tools. This understanding can lead to more user-centred and culturally responsive platform designs. The generalizability of the findings from Bilibili to other platforms remains to be explored; other Chinese social media platforms such as Sina Weibo and WeChat have their own distinctive cultures of use, including as regards graphicons. Such explorations can benefit from the historically informed socio-technical approach, which sheds light on the evolution of communication tools and digital communication, as well as broader trends in digital culture.

Acknowledgements

This study was partially supported by the Humanities and Social Sciences Fund of the Ministry of Education of the People's Republic of China (Grant No. 22YJA740037) and the Department of Education of Guangdong Province (Grant No. 2020WTSCX016). We also acknowledge the generous contributions of the China Scholarship Council, the Carnegie Corporation of New York, the Ford Foundation, the Harvard-Yenching Institute, the Henry Luce Foundation, and the Rockefeller Brothers Fund. The views and opinions expressed in the article do not reflect the positions of the supporting institutions.

About the authors

Yiqiong Zhang is Professor of Applied Linguistics at the Centre for Linguistics & Applied Linguistics at Guangdong University of Foreign Studies in China. She obtained her PhD from the National University of Singapore in 2013. Her research focuses mainly on digital communication, covering areas such as Multimodal Discourse Analysis, Science Communication, Cross-cultural Studies, and Multiliteracies. She is on the Editorial Board of the journals Discourse & Society, Multimodality & Society, and Computers and Composition. She can be contacted at zhangyiqiong@gdufs.edu.cn

Susan C. Herring is Professor of Information Science and Linguistics and Director of the Centre for Computer-Mediated Communication at Indiana University, Bloomington. She received her PhD in Linguistics from the University of California, Berkeley in 1991. One of the first scholars to apply linguistic methods of analysis to computer-mediated communication (CMC), she created the Computer-Mediated Discourse Analysis (CMDA) paradigm. Her recent research has focused on multimodal CMC, including communication mediated by telepresence robots, graphical icons, and augmented reality video filters. She is a past editor of the Journal of Computer-Mediated Communication and currently edits the online journal Language@Internet. She can be contacted at herring@iu.edu

Yanmin Wu is a Lecturer at the School of Foreign Languages, Guangzhou College of Applied Science and Technology. She holds an MA in applied linguistics from Guangdong University of Foreign Studies. Her research interest includes social media studies and multimodal discourse analysis. She can be contacted at yanminwu2020@163.com

Rongle Tan is a PhD candidate at the School of Education, Macquarie University. His research focuses on digital communication, educational technologies, multimodal discourse analysis, and social semiotics. His recent publications appear in Learning, Media and Technology, Discourse Studies, and Discourse & Communication. He can be contacted at rongle.tan@hdr.mq.edu.au.

References

Ahn, W., Park, J., & Han, K. (2011). Emoticons convey emotion in CMC. In L. Little & L. Coventry (Eds.), Proceedings of HCI 2011: The 25th BCS Conference on Human Computer Interaction (HCI). BCS Learning & Development Ltd. https://doi.org/10.14236/ewic/HCI2011.11

Albert, G. (2020). Beyond the binary: emoji as a challenge to the image-word distinction. In C. Thurlow, C. Dürscheid, & F. Diémoz (Eds.), Visualizing digital discourse (pp. 65-80). De Gruyter Mouton.

Berard, B. (2018). I second that emoji: the standards, structures, and social production of emoji. First Monday, 23(9). https://doi.org/10.5210/fm.v23i9.9381

Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: a socio-technical perspective. Part I: The causes. MIS Quarterly, 1(3), 17-32. https://doi.org/10.2307/248710

Chang, C. Y. (2016). EFL reviewers' emoticon use in asynchronous computerm-ediated peer response. Computers and Composition, 40, 1-18. https://doi.org/https://doi.org/10.1016/j.compcom.2016.03.008

Chik, A., & Vásquez, C. (2017). A comparative multimodal analysis of restaurant reviews from two geographical contexts. Visual Communication, 16(1), 3-26. https://doi.org/10.1177/1470357216634005

Danet, B. (2001). Cyberplay: communicating online. Berg.

de Seta, G. (2018). Biaoqing: the circulation of emoticons, emoji, stickers, and custom images on Chinese digital media platforms. First Monday, 23(9). https://doi.org/10.5210/fm.v23i9.9391

Gao, J. (2021). Business model evolution of content platform in emerging economy: a case study of Bilibili. American Journal of Industrial and Business Management, 11(8), 954-972. https://doi.org/10.4236/ajibm.2021.118058

Ge, J. (2020). The anatomy of memetic stickers: An analysis of sticker competition on Chinese social media. In Proceedings of the 3rd International Workshop on Emoji Understanding and Applications in Social Media (Emoji2020). DOI: 10.36190/2020.01

Ge-Stadnyk, J., & Herring, S. C., & Sa, L. (2025). Multimodality in digital discourse: exploring image-text interrelations in WeChat sticker use. In C. Vasquez & J. Chovanec (Eds.), Experiencing digital discourses: multimodality, engagement, activism. Palgrave Macmillan.

Grundlingh, L. (2018). Memes as speech acts. Social Semiotics, 28(2), 147-168. https://doi.org/10.1080/10350330.2017.1303020

Guntuku, S. C., Li, M., Tay, L., & Ungar, L. H. (2019). Studying cultural differences in emoji usage across the East and the West. International AAAI Conference on Web and Social Media, Münich, Germany.

Hanchard, M. (2024). Engaging with digital maps: our knowledgeable deferral to rough guides. Palgrave Macmillan. https://doi.org/10.1007/978-981-99-8972-0

Herring, S. C. (2004). Computer-mediated discourse analysis: an approach to researching online behavior. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), Designing for virtual communities in the service of learning (pp. 338-376). Cambridge University Press.

Herring, S. C., & Dainas, A. (2017). "Nice picture comment!": graphicons in Facebook comment threads. Proceedings of the Fiftieth Hawaii International Conference on System Sciences (pp. 2185-2194). IEEE Press. https://doi.org/https://doi.org/10.24251/HICSS.2017.264

Herring, S. C., & Ge-Stadnyk, J. (2024). Emoji and illutionarity: acting on, and acting as, language. In M. Gill, A. Malmivirta, & B. Warvik (Eds.), Structures in discourse: studies in interaction, adaptability, and pragmatic functions (pp. 124-165). John Benjamins.

Kaneyasu, M. (2022). Multimodal strategies for balancing formality and informality: the role of kaomoji in online comment-reply interactions. Internet Pragmatics, 5(1), 143-164. https://doi.org/https://doi.org/10.1075/ip.00071.kan

Kling, R. (2007). What is social informatics and why does it matter? The Information Society, 23(4), 205-220. https://doi.org/10.1080/01972240701441556

Koch, T. K., Romero, P., & Stachl, C. (2022). Age and gender in language, emoji, and emoticon usage in instant messages. Computers in Human Behavior, 126, 106990. https://doi.org/https://doi.org/10.1016/j.chb.2021.106990

Konrad, A., Herring, S. C., & Choi, D. (2020). Sticker and emoji use in Facebook messenger: Implications for graphicon change. Journal of Computer-Mediated Communication, 25(3), 217-235. https://doi.org/https://doi.org/10.1093/jcmc/zmaa003

Kozar, S. (1995). Enduring traditions, ethereal transmissions: recreating Chinese New Year celebrations on the Internet. Journal of Computer-Mediated Communication, 1(2). https://doi.org/https://doi.org/10.1111/j.1083-6101.1995.tb00329.x

Li, W., & Zhu, H. (2019). Tranßcripting: playful subversion with Chinese characters. International Journal of Multilingualism, 16(2), 145-161. https://doi.org/https://doi.org/10.1080/14790718.2019.1575834

Liu, C., Tan, X., Zhou, T., Zhang, W., Liu, J., & Lu, X. (2022). Emoji use in China: popularity patterns and changes due to COVID-19. Applied Intelligence, 52(14), 16138-16148. https://doi.org/10.1007/s10489-022-03195-y

Liu, F., Wang, H., & Xu, X. (2020). Usage behaviour and cause of emoji in social media. Complex Systems and Complexity Science, 17(3), 70-77. https://doi.org/10.13306/j.1672-3813.2020.03.007

Logi, L., & Zappavigna, M. (2021). A social semiotic perspective on emoji: how emoji and language interact to make meaning in digital messages. New Media & Society, 25(12), 3222-3246. https://doi.org/10.1177/14614448211032965

Ma, X. (2016). From Internet memes to emoticon engineering: insights from the Baozou comic phenomenon in China. In M. Kurosu (Ed.), Human-computer interaction. Novel user experiences: 18th International Conference, HCI International 2016 (pp. 15-27). Springer.

Markman, K. M., & Oshima, S. (2007). Pragmatic play? some possible functions of English emoticons and Japanese kaomoji in computer-mediated discourse. Paper presented at the Association of Internet Researchers Annual Conference 8.0: Let's Play!, Vancouver, B.C., Canada.

Morgan, C. (2022). Current digital archaeology. Annual Review of Anthropology, 51, 213-231. https://doi.org/10.1146/annurev-anthro-041320-114101

Pavalanathan, U., & Eisenstein, J. (2016). More emojis, less :) The competition for paralinguistic function in microblog writing. First Monday, 22(11). https://doi.org/https://doi.org/10.5210/fm.v21i11.6879

Pohl, H., Domin, C., & Rohs, M. (2017). Beyond just text: semantic emoji similarity modeling to support expressive communication. ACM Transactions on Computer-Human Interaction (TOCHI), 24(1), 1-42. https://doi.org/10.1145/3039685

Provine, R. R., Spencer, R. J., & Mandell, D. L. (2007). Emotional expression online: emoticons punctuate website text messages. Journal of Language and Social Psychology, 26(3), 299-307. https://doi.org/https://doi.org/10.1177/0261927X06303481

Qin, Y., Shen, G., Zhao, W., Chen, Y., Yu, M., & Jin, X. (2019). A network security entity recognition method based on feature template and CNN-BiLSTM-CRF. Frontiers of Information Technology & Electronic Engineering, 20(6), 872-884. https://doi.org/https://doi.org/10.1631/FITEE.1800520

Sandel, T. L., Buttny, R., & Varghese, M. (2019). Online interaction across three contexts: an analysis of culture and technological affordances. Journal of Intercultural Communication Research, 48(1), 52-71. https://doi.org/10.1080/17475759.2018.1552616

Sun, L., Zhang, J., & Shou, J. (2023). Value co-creation of community e-commerce platform: how does Bilibili balance between content and commercialization? In J. Zhang, K. Ying, K. Wang, Z. Fan, & Z. Zhao (Eds.), Innovation of digital economy: Cases from China (pp. 117-128). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-1741-9_8

Tagg, C., & Lyons, A. (2021). Repertoires on the move: exploiting technological affordances and contexts in mobile messaging interactions. International Journal of Multilingualism, 18(2), 244-266. https://doi.org/10.1080/14790718.2020.1867150

Tang, Y., & Hew, K. F. (2018). Emoticon, emoji, and sticker use in computer-mediated communications: understanding its communicative function, impact, user behavior, and motive. In L. Deng, W. W. K. Ma, & C. W. R. Fong (Eds.), New media for educational change (pp. 191-201). Springer.

Tang, Y., & Hew, K. F. (2019). Emoticon, emoji and sticker use in computer-mediated communication: A review of theories and research findings. International Journal of Communication, 13, 27.

Tian, H. (2021). Analysis of the marketing strategy of Bilibili and the reasons for its success. In R. Thurasamy, I. A. Majid, N. A. B. A. Rahman, & A. H. Zulkafli (Eds.), Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021) (pp. 2853–2856). Atlantis Press. https://doi.org/10.2991/assehr.k.211209.463

Trist, E. L. (1978). On socio-technical systems. In W. A. Pasmore & J. J. Sherwood (Eds.), Sociotechnical systems: a sourcebook (pp. 43-57). University Associates.

van Dijck, J. (2013). The culture of connectivity: a critical history of social media. Oxford University Press.

Wang, Y., Zhang, Y., Zhang, G., He, S., & Qi, J. (2024). Linguistics properties of emojis: a quantitative exploration of emoji frequency, category, and position on Twitter. Journal of Quantitative Linguistics, 31(3), 183-209. https://doi.org/https://doi.org/10.1080/09296174.2024.2347055

Wu, Q., Sang, Y., Zhang, S., & Huang, Y. (2018). Danmaku vs. forum comments: understanding user participation and knowledge sharing in online videos. In A. Forte, M. Prilla, & A. Vivacqua (Eds.), Proceedings of the 2018 ACM International Conference on Supporting Group Work (pp. 209–218). Association for Computing Machinery. https://doi.org/10.1145/3148330.3148344

Yang, G. (2021). Social media and state-sponsored platformization in China. In G. Yang & W. Wang (Eds.), Engaging social media in China: platforms, publics, and production (pp. xi-xxxi). Michigan State University Press.

Zhang, Y., Herring, S. C., & Gan, S. (2022). Graphicon evolution on the Chinese social media platform Bilibili. In S. Wijeratne, J. Lee, H. Saggion, & A. Sheth (Eds.), Proceedings of the Fifth International Workshop on Emoji Understanding and Applications in Social Media (Emoji2022) (pp. 75–85). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emoji-1.9

Zhang, Y., Wang, M., & Li, Y. (2021). More than playfulness: emojis in the comments of a WeChat official account. Internet Pragmatics, 4(2), 247-271. https://doi.org/https://doi.org/10.1075/ip.00048.zha

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Analyzing the language of rejection: a study of user flagging responses to hate speech on Reddit

Sharon Lisseth Perez, Xiaoying Song, and Lingzi Hong

DOI: https://doi.org/10.47989/ir30iConf47191

Abstract

Introduction. Online hate speech poses significant threats to individuals and society, exacerbating psychological harm, discrimination, and potential real-world violence. While automated detection models are available, their inability to recognize subtle variations of hate speech, particularly implicit forms, emphasizes the need for supplementary methods.

Method. This study investigates the potential of user-written flagging messages in enhancing hate speech detection, focusing on the characteristics and identification of flagging messages. We created a dataset of flagging messages and the comments they respond to, employing transformer-based models (BERT, RoBERTa, ALBERT, DistilBERT, and XLNet) for classification.

Analysis. Linguistic analysis using SEANCE and Named Entity Recognition was conducted to reveal unique characteristics of flagging messages.

Results. Our findings show that BERT and DistilBERT models achieved the highest accuracy in classifying flagging messages, with distinct linguistic patterns emerging in flagging content.

Conclusion. This research contributes to the development of more nuanced hate speech detection methods by leveraging user-generated flagging content. These findings have implications for improving automated content moderation systems and supporting more inclusive online environments. Future work will focus on the effectiveness of flagging messages in identifying implicit hate speech across diverse cultural contexts.

Introduction

Online hate speech poses significant threats to individuals and communities, exacerbating psychological harm, discrimination, and potential real-world violence. Extensive research has been dedicated to identifying and addressing this issue, with various studies utilizing automated methods, including natural language processing techniques, to detect and categorize hate speech on social media platforms (Salminen et al., 2020, Yu et al., 2022).

However, current detection methods primarily focus on explicit forms of hate, often overlooking more subtle manifestations and evolving abusive language. This limitation highlights the need for more nuanced approaches that can capture the complex dynamics of online hate. One potential source of insight that has been under-explored is user-written flagging messages.

Flagging is widely employed by users across social media platforms to report offensive content, as shown in Figure 1, and it provides moderators with a rhetorical defence for content removal decisions (Crawford and Gillespie, 2016). In an environment where control and transparency are limited, flags are crucial in giving users a voice (Zhang et al., 2023). Chandrasekharan et al. (2018) found that online posts flagged by regular users and later reviewed by moderators were important in determining the best ways to intervene.

This study aims to identify flagging messages, laying a foundation for the empirical investigation of flagging messages in enhancing hate speech detection, particularly for subtle forms of abuse or microaggressions (De la Peña Sarracén and Rosso, 2023; MacAvaney et al., 2019). Specifically, we address the following research questions:

1. What are the linguistic features that distinguish flagging from non-flagging messages?
2. What NLP models are most effective in accurately detecting flagged content?
3. To what extent can flagging messages aid in identifying implicit hate, and what insights can they provide into this form of hate expression?

We aim to provide insights that can inform the development of more comprehensive and nuanced hate speech detection models, ultimately contributing to safer and more inclusive online environments.

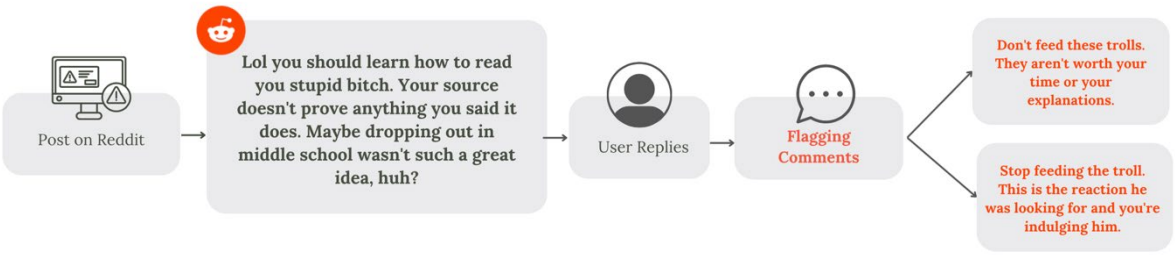


Figure 1. Online users post flagging messages to warn others. These comments indicate the comment they respond to is harmful speech

Related work

Implicit hate speech detection

Detecting implicit forms of hate speech poses significant challenges due to linguistic nuances and contextual dependencies. ElSherief et al. (2021) presented a comprehensive benchmark dataset aimed at understanding and detecting implicit hate speech, providing a foundation for evaluating detection algorithms. Ghosh et al. (2023) proposed CoSyn, a context-synergized hyperbolic network that considers both personal and dialogue context in conversation trees. Their approach highlights the importance of contextual information in this task. Jafari et al. (2023) investigated the

influence of fine-grained emotions on implicit hate speech classification. Their research employed single task learning and multi-task learning models and revealed that integrating emotional features improved hate speech detection. Recently, Ahn et al. (2024) introduced a SharedCon model that leverages shared semantic information to enhance the detection of implicit hate speech.

These works collectively advance the understanding of implicit hate speech. However, as hate language evolves, the curated implicit hate speech datasets could slowly become obsolete, affecting detection accuracy. We propose a detection method that is anchored in flagging messages, which are less likely to change due to the lack of motivation for alteration among users.

Method

Data collection and processing

We started by compiling a small selection of comments from Reddit discussion threads. We collected a total of 1,050 pairs of hate speech comments and their corresponding replies from 39 subreddits that have been identified as hateful (Vidgen et al., 2021). Examples of these subreddits include r/TwoXChromosomes, r/TrueReddit, and r/ChangeMyView. We used Reddit's API to collect the comments while ensuring that we followed the platform's terms of service.

Our research team manually examined and labeled each reply message. We utilized a binary (0,1) coding system to categorize the messages based on their content, purpose, tone, effectiveness, and rejection in addressing hate speech. Inter-rater agreement was calculated to ensure the reliability of the labeling process. The obtained results are as follows: Cohen's kappa: 0.81 and Krippendorff's alpha: 0.82. A Kappa coefficient between 0.61 and 0.80 indicates substantial agreement (Viera et al., 2005), demonstrating the reliability of our labels.

Data analysis

We utilize the SEANCE tool (Sentiment Analysis and Cognition Engine) to perform in-depth sentiment, social cognition, and social order analysis on textual data. Its comprehensive set of indices and component scores allows for a nuanced and insightful analysis of language use, emotions, and social dynamics (Crossley et al., 2017). In addition to the SEANCE analysis, we conducted a Named Entity Recognition (NER) using Spacy (Naseer et al., 2021).

We use the Wilcoxon Rank Sum Test to conduct statistical tests on SEANCE and NER features between the flagging and non-flagging messages. Then Bonferroni correction is conducted to identify the most significant features.

Classification experiments

We utilize several state-of-the-art transformer-based models, including BERT (Kenton and Toutanova, 2019), RoBERTa (Liu, 2019), ALBERT (Lan, 2019), DistilBERT (Sanh, 2019), and XLNet (Yang, 2019) to classify a comment to be flagging or not.

The labeled data was split into training and testing sets using an 8:2 ratio, ensuring class balance by stratifying based on the 'label.' As the labels of 'flagging' and 'non-flagging' are unbalanced, we experiment with sampling and data augmentation methods to investigate whether these techniques will benefit the prediction performance. We first use undersampling, which involves reducing the number of records in the majority class. Random undersampling has been empirically shown to be one of the most effective resampling methods. Few of the more sophisticated undersampling methods have outperformed random undersampling in empirical studies (Liu, 2004).

We then expand our dataset to evaluate whether implementing NLTK tools such as punkt, averaged_perceptron_tagger, and wordnet would enhance our models' performance. Punkt is a tokenizer that uses an unsupervised algorithm to divide a text into a list of sentences, taking into

consideration abbreviation words, collocations, and words that start sentences (Natural Language Toolkit, n.d.).

The models were trained using the PyTorch framework and the Hugging Face Transformers library. Model evaluation metrics included accuracy, precision, recall, and F1-score. We generated classification reports to provide detailed insights into model performance across different classes.

Results

In this section, we present detailed findings of our experiments, encompassing both the insights gained from linguistic analysis techniques and the performance of various transformer-based models.

Linguistic analysis

This analysis offers insights into the linguistic characteristics that differentiate flagging from non-flagging messages, providing a deeper understanding of their nature.

Our SEANCE analysis revealed several significant linguistic differences between flagging and non-flagging comments, with a particularly strong emphasis on negative emotion words in flagging content. Statistically significant differences ($p < 0.001$) were observed in various measures. For example, anger (Lexicon) was considerably more prevalent in flagging comments, with a mean of 0.073 compared to 0.029 in non-flagging comments. Fear and disgust showed the largest difference, with means of 0.437 and 0.210 for flagging and non-flagging comments respectively. General negative emotions (Negative_EmoLex) were also more common in flagging comments (mean 0.104) compared to non-flagging ones (mean 0.052).

These findings indicate that flagging comments are characterized by a significantly higher emotional intensity, specifically negative emotions like anger, fear, and disgust. This is consistent with prior research on hate speech and offensive language, which frequently involve strong expressions of negative emotions (Yu et al., 2022). We also observed differences in other linguistic features. For example, the second-person pronouns (You_GI) were higher in flagging comments, potentially indicating more direct confrontational language. Table 1 shows these findings.

Variable Category	Variable	Mean (Flagging)	Mean (Non-flagging)	p-Value	Bonferroni Correction
Negative Emotion words	Anger (Lexicon)	0.073	0.029	$2.47E^{-41}$	$6.66E^{-39}$
	Fear and Disgust	0.437	0.210	$6.30E^{-39}$	$1.70E^{-36}$
	Negative Emotions (Negative_EmoLex)	0.104	0.052	$6.81E^{-38}$	$1.83E^{-35}$
	Fear (Lexicon)	0.072	0.029	$7.45E^{-36}$	$2.01E^{-33}$
	Negative Adjectives	0.733	0.364	$6.08E^{-10}$	$1.64E^{-07}$
	Negative Words	0.063	0.043	$8.39E^{-10}$	$2.27E^{-07}$
	Negative Sentiment (VADER)	0.144	0.105	$3.71E^{-09}$	$1.00E^{-06}$
	Hatred (Lexicon)	0.002	0.000	$1.93E^{-05}$	0.005
	Disgust (Lexicon)	0.027	0.020	$1.01E^{-04}$	0.027
	Negativity (Lexicon)	0.074	0.062	$3.19E^{-04}$	0.086
Reference	Second-Person Pronouns (You_GI)	0.041	0.031	$1.78E^{-06}$	0.000
Quality and quantity	Numerical Mentions	0.004	0.009	$9.64E^{-05}$	0.026
	Frequency References	0.010	0.005	$3.06E^{-04}$	0.083
Other affect	Sentiment Compound (VADER)	-0.113	-0.005	$1.02E^{-08}$	$2.75E^{-06}$
Emotion words	Positive Nouns	-0.177	-0.088	$1.90E^{-07}$	0.000

Table 1. The most significant linguistic differences are grouped by variable category. Only variables that pass the Bonferroni correction are included

NER analysis

This analysis assists in uncovering contextual elements and specific entity types that might indicate potentially problematic content. The results show a statistically significant difference between PERSON and CARDINAL. The presence of person names (PERSON) is strongly associated with non-flagging comments. This suggests that comments mentioning specific individuals are less likely to be a flagging message. The higher presence of cardinal numbers (CARDINAL) in non-flagging comments indicates that more factual or quantitative content is less likely to be the flagging message. These results are displayed in Table 2.

GPE (Geo-Political Entities) tends to appear more often in non-flagging comments. The trend with GPEs, while not significant after correction, hints that mentions of geographical or political entities might be more common in non-flagging comments.

Variable	Mean (Flagging)	Mean (Non-flagging)	p-value	Bonferroni correction
PERSON	0.039	0.095	0.000	0.002
CARDINAL	0.043	0.080	0.001	0.011
GPE	0.038	0.064	0.015	0.212

Table 2. Entities with more prevalent differences. Only variables that pass the Bonferroni correction are included

Model performance

We examine the effectiveness of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet models. We conduct experiments using various input configurations, including replies to hate speech alone (Child), and a combination of hate speech (Parent) and replies to it. Table 3 presents these results.

BERT and DistilBERT show the highest overall performance, both achieving an accuracy of 0.81 and balanced F1 scores across labels. BERT, using only Child, demonstrates the highest F1 score (0.82) for flagging comments.

There is a slight variation in how models handle different contexts (only Child vs. Parent and Child). Most models demonstrate balanced performance between labels of flagging and non-flagging. Including Parent as a context varies across models: BERT and XLNet perform best with only Child, suggesting they may be more sensitive to noise in Parent comments. In contrast, RoBERTa, ALBERT, and DistilBERT show strong performance with both Parent and Child, indicating better context integration capabilities.

DistilBERT's high performance is particularly noteworthy, as it is a compressed model designed for efficiency. The similar performance of BERT and DistilBERT suggests that DistilBERT might be more suitable in resource-constrained environments.

Model	***Input***	**Non-Flagging**			**Flagging**			**Weighted Average**			***Accuracy***
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
BERT	Child only	0.84	0.76	0.80	0.78	0.85	**0.82**	0.81	0.81	**0.81**	**0.81**
RoBERTa	Parent and child	0.77	0.81	0.79	0.80	0.75	0.77	0.78	0.78	0.78	0.78
ALBERT	Parent and child	0.82	0.77	0.79	0.78	0.83	0.80	0.80	0.80	0.80	0.80
DistilBERT	Parent and child	0.82	0.80	0.81	0.80	0.82	0.81	0.81	0.81	**0.81**	**0.81**
XLNet	Child only	0.80	0.77	0.79	0.78	0.81	0.79	0.79	0.79	0.79	0.79

Table 3. Summary of best-performing model results, including the impact of different input configurations (Child only or Parent and child) on the model's performance

Conclusion

Our study provides insights into the linguistic characteristics of flagging comments. Our study combining Linguistic and NER analysis, and transformer-based models, reveals several key findings:

The Linguistic analysis demonstrated a strong association between negative emotional language and flagging comments. The NER analysis revealed that certain types of named entities, particularly PERSON and CARDINAL, are differentially associated with flagging and non-flagging comments. These findings could be instrumental in identifying flagging messages, which could be used to develop more sophisticated content moderation tools. The evaluation of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet models demonstrated their effectiveness in detecting flagging messages. The context should be carefully considered when designing flagging systems.

Limitations and future directions

We have identified flagging messages; how these could enhance the detection of implicit or varied forms of hate speech will be further investigated in the next step, as the work is still in progress. While our study gives us valuable insights, it's important to recognize its limitations. The binary classification of comments as flagging or not may simplify the complex nature of online discourse. Future research will explore more granular categorizations, considering the nuances of different types and intensities of negative emotions and their relationship to various forms of online harm

like implicit hate. The trending large language models may have better performance in identifying flagging messages, which will be experimented in the future.

This study contributes to the ongoing efforts to create safer and healthier online discussions and communities. Yet, the scope of our research can be expanded and address a broader range of online harm indicators.

Acknowledgements

The authors gratefully acknowledge the financial support from the Institute of Museum and Library Services under Grant LG-256661-OLS-24 and LG-256666-OLS-24. Sharon Lisseth Perez also acknowledges the support of the Fulbright Program for funding her PhD studies at the University of North Texas. The authors would like to thank the collaborators who contributed to this research.

About the authors

Sharon Lisseth Perez is a PhD student in the College of Information at the University of North Texas, Texas, USA. Her research interests are on natural language processing, artificial intelligence, and their applications in education and social justice. She can be reached via email at sharonperez@my.unt.edu

Xiaoying Song is a PhD student in College of Information, University of North Texas, Texas, USA. Her research focuses on evaluating generative counter-speech and addressing health misinformation. She can be reached via email at xiaoyingsong@my.unt.edu

Lingzi Hong is an Assistant Professor of Data Science in the College of Information at the University of North Texas. She holds a Ph.D. in Information Science from the University of Maryland, College Park. Her research focuses on human-centered computing and artificial intelligence. She can be reached at lingzi.hong@unt.edu

References

Ahn, H., Kim, Y., Kim, J., & Han, Y. S. (2024, August). SharedCon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics ACL 2024* (pp. 10444-10455). https://aclanthology.org/2024.findings-acl.622/

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., ... & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1-25. https://doi.org/10.1145/3274301

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410-428. https://doi.org/10.1177/1461444814543163

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49, 803-821. https://doi.org/10.3758/s13428-016-0743-z

De la Peña Sarracén, G. L., & Rosso, P. (2023). Systematic keyword and bias analyses in hate speech detection. *Information Processing & Management*, *60*(5), 103433. https://doi.org/10.1016/j.ipm.2023.103433

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021, November). Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 345-363). https://aclanthology.org/2021.emnlp-main.29/

Ghosh, S., Suri, M., Chiniya, P., Tyagi, U., Kumar, S., & Manocha, D. (2023). CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network. *arXiv preprint arXiv*:2303.03387. https://doi.org/10.48550/arXiv.2303.03387

Jafari, A. R., Li, G., Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2023). Fine-grained emotions influence on implicit hate speech detection. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3318863

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, p. 2). https://doi.org/10.48550/arXiv.1810.04805

Lan, Z. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942

Liu, A. Y. C. (2004). The effect of oversampling and undersampling on classifying imbalanced text datasets. http://dx.doi.org/10.26153/tsw/12300

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv*:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Naseer, S., Ghafoor, M. M., bin Khalid Alvi, S., Kiran, A., Rahmand, S. U., Murtazae, G., & Murtaza, G. (2021). Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance. *Pakistan Journal of Multidisciplinary Research*, 2(2), 293-308. https://pjmr.org/pjmr/article/view/150

Natural Language Toolkit. (n.d.). *NLTK 3.0 documentation*. https://www.nltk.org/index.html

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10, 1-34. https://doi.org/10.1186/s13673-019-0205-6

Sanh, V. (2019). DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., & Tromble, R. (2021). Introducing CAD: the contextual abuse dataset. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2289-2303) https://doi.org/10.18653/v1/2021.naacl-main.182

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5), 360-363. https://api.semanticscholar.org/CorpusID:38150955

Yang, Z. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237. https://doi.org/10.48550/arXiv.1906.08237

Yu, X., Blanco, E., & Hong, L. (2022). Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In Proceedings of the 2022 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies (pp. 5918-5930). https://doi.org/10.48550/arXiv.2206.06423

Zhang, A. Q., Montague, K., & Jhaver, S. (2023). Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Posts. *arXiv preprint arXiv*:2309.06688. https://doi.org/10.48550/arXiv.2309.06688

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Comunicação científica e mediação da informação em mídia social: relato de experiência dos eventos organizados pela Comissão Diálogos de Pesquisa do PPGCI/UFF

Camilla Castro de Almeida, Juliana Maia Mendes, Vinicius Ribeiro Soares dos Santos, Elisabete Gonçalves de Souza, e Lucia Maria Velloso de Oliveira

DOI: https://doi.org/10.47989/ir30iConf47188

Abstract

Introduction. '*Research dialogues*' is a scientific dissemination event developed by a committee associated with the Information Science Postgraduate Program of the Universidade Federal Fluminense (Rio de Janeiro, Brazil). As a lecture series, it allows the exchange of knowledge among a specialized community, through the presentation of doctoral and post-doctoral reports, however, it is not restricted to this, seeking to get closer to undergraduate students. This study aims to outline the activities carried out by the committee between 2020 and 2024, covering 20 editions of the event.

Method. It uses the theoretical framework of mediation to analyze and interpret the results regarding the audience profile and the evaluation of these editions. The methodology adopts a descriptive, qualitative-quantitative approach typical of an experience report.

Results. The results indicate that the Committee organizes events that focus on two main categories: 1) post-doctoral internship closing seminars and 2) doctoral research meetings. These events effectively disseminate scientific knowledge and foster closer ties between postgraduate and undergraduate studies.

Conclusion. They contribute significantly to scientific literacy by facilitating direct mediation through dialogue, questioning, and reflection among presenters, facilitators, and audiences.

Introdução

A gênese das ciências documentais está na organização e preservação dos documentos, mas esse quadro vem mudando ao longo dos séculos com o desenvolvimento de tecnologias como a imprensa, que deu origem aos catálogos impressos, listas de inventários etc. Com o advento das tecnologias digitais e o surgimento das bases de dados de acesso remoto, a questão da disseminação da informação tornou-se predominante, surgindo novas possibilidades para a difusão do conhecimento preservado, para alcance de diferentes públicos.

Conforme ressaltam Brito e Valls (2015, p. 4), trata-se de um nova perspectiva de democratização do acesso que se expande para além dos documentos tradicionais e inclui "*outras mídias em diferentes formatos, que podem ser acessadas a qualquer hora em qualquer lugar pela Internet*". Muda-se a relação espaço-tempo nos processos de mediação da informação, o que leva ao surgimento de um novo cenário de práticas que facilitam a disseminação da informação e do conhecimento viabilizados pelas tecnologias das mídias sociais.

Nos estudos da Ciência da Informação (CI), a mediação assume um papel fundamental, sendo analisada sob a perspectiva da interação da informação com o usuário, podendo ser realizada de forma indireta, por meio da elaboração de instrumentos de disseminação da informação, ou de forma direta, através do oferecimento de serviços de consulta ou pela realização de ações acadêmicas e culturais (Pimenta; Almeida Júnior, 2009). Um dos objetivos das ações dessa natureza é a publicização do conhecimento e o letramento científico (Cunha, 2017). Cunha (2017) categoriza os eventos acadêmico-científicos de difusão de um conhecimento específico, como "*abordagem cívica*", no sentido de serem pensados para atingir não exclusivamente uma comunidade de especialistas, mas incluir outros segmentos sociais.

Por parte de instituições de pesquisa, como as universidades, as tecnologias digitais de comunicação trouxeram inovação para o campo da comunicação científica. Além de usarem canais tradicionais para a difusão de suas pesquisas, passaram a usar as mídias sociais para divulgar e compartilhar conteúdos, criando eventos específicos para tais fins, além de articular estas atividades a práticas tradicionais e inovadoras de mediação como a presença de moderadores e espaços para interlocução com os participantes por meio do *chat* da plataforma. Nesse contexto, o que as diferenciam das mídias tradicionais é o alcance e o registro imediato para disponibilização do conteúdo ao público interessado, aspecto relevante no sentido que proporciona novas modalidades de mediação e difusão de conhecimentos. Outro aspecto a ser destacado é a disponibilização imediata do conteúdo, permitindo a visualização do evento por aqueles que não puderam assistir ao vivo.

No âmbito desse artigo, nos dedicamos a relatar e analisar uma ação de mediação de difusão de conhecimento acadêmico-científico proporcionada pela realização sistemática de eventos de divulgação das pesquisas realizadas pelos doutorandos, egressos e pesquisadores de pós-doutorado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal Fluminense (PPGCI/UFF). A série de palestras foi intitulada *Diálogos de Pesquisa*.

A série *Diálogos de Pesquisa* é um evento de compartilhamento de conhecimentos entre uma comunidade científica especializada: os pesquisadores, doutores e pós-doutores em Ciência da Informação. As palestras são realizadas em formato eletrônico, via *YouTube* (*lives*), no âmbito dos estudos doutorais e pós-doutorais desenvolvidos no Programa. É, portanto, uma forma de disseminar o conhecimento produzido pelos discentes egressos a um público potencialmente maior do que seria alcançado em formato presencial.

O trabalho objetiva: (1) relatar os eventos desenvolvidos no âmbito desta série de comunicações científicas, considerando como aporte teórico o conceito de mediação; (2) conhecer e analisar o perfil do público; e (3) avaliar sua frequência, interação e grau de satisfação.

Comunicação científica e mediação

Segundo Le Coadic (1996 *apud* Targino, 2000, p. 9-10),

> *As atividades científicas e técnicas são o manancial de onde surgem os conhecimentos científicos e técnicos que se transformarão, depois de registrados, em informações científicas e técnicas. [...] Sem informação a pesquisa seria inútil e não existiria o conhecimento. Fluido precioso, continuamente produzido e renovado, a informação só interessa se circula, e, sobretudo, se circula livremente.*

No âmbito dos estudos sobre informação científica, vale distinguir dois conceitos: comunicação científica e divulgação científica. Enquanto a comunicação científica visa à disseminação de informação especializada entre pares a fim de tornar públicos à comunidade científica os avanços obtidos no campo científico, a divulgação científica objetiva o letramento científico, democratiza o acesso da sociedade à informação científica (Bueno, 2011).

A formalização da comunicação científica é resultado da necessidade de troca de informações e do compartilhamento dos resultados das pesquisas entre pares, visto que seu exercício de fazer ciência, por vezes solitário, também é significativamente social. Usualmente, no campo científico, o processo de comunicação entre os pesquisadores (e entre pesquisadores e o público) acontece por meio de dois canais que se complementam: os canais de comunicação formal e de comunicação informal.

Os canais de comunicação são tradicionalmente caracterizados conforme alguns aspectos, tais como: público, armazenamento, rapidez e direção do fluxo da informação, redundância da informação e formas de avaliação (Quadro 1).

Canais formais	Canais informais
Público potencialmente grande	Público restrito
Informação armazenada e recuperável	Informação não armazenada e não recuperável
Informação relativamente antiga	Informação recente
Direção do fluxo selecionada pelo usuário	Direção do fluxo selecionada pelo produtor
Redundância moderada	Redundância, às vezes, significativa
Avaliação prévia	Sem avaliação prévia
Feedback irrisório para o autor	*Feedback* significativo para o autor

Quadro 1. Características tradicionais de canais formais e informais de comunicação científica (Fonte: Targino, 2000, p. 19)

Com base nas diferenças elencadas, alguns exemplos de canais informais são as reuniões e trocas de e-mail entre pesquisadores, enquanto como canais formais podemos citar publicações técnicas, livros, periódicos científicos, congressos, etc. A mediação da informação desses materiais realiza-se por meio da intervenção do profissional que descreve e representa os conteúdos e os disponibiliza em catálogos e outras ferramentas de pesquisa, como bibliografias, lista de inventário, etc., ou por meio da realização de transmissões de vídeo, recurso de mediação utilizado pela *Diálogos de Pesquisa*.

É fundamental pontuar que, com a popularização da comunicação digital, novas ferramentas tecnológicas naturalmente passaram a impactar a comunicação entre cientistas, o que suaviza a dicotomia apresentada no modelo de Targino (2000) sobre canais de comunicação científica, resultando em um modelo híbrido. Costa (2009) ressalta que no modo híbrido de comunicação os canais informais predominam sobre os formais, sendo o uso de recursos eletrônicos o principal meio de produção, organização e difusão do conhecimento científico.

Contudo, para pensarmos sobre a mediação da informação, é necessário irmos para além dos suportes, da relação emissor/receptor e dos sistemas de armazenamento e de busca; também é

necessário refletir sobre os contextos da produção da informação que envolvem emissão, difusão, circulação e recepção (Feitosa, 2016), aspecto ao qual atemo-nos ao analisar a dinâmica do evento e o perfil dos participantes.

Seguindo a classificação tradicional dos canais de comunicação científica, as palestras podem ser classificadas como comunicação via canal informal, já que fazem uso da fala e da audição como recursos principais da comunicação, sendo uma troca direta entre as pessoas, em nosso caso, sem produção de publicações como livros, anais de eventos ou periódicos. Nessa direção, categorizamos as palestras como ações de mediação direta, ou seja, aquela que exige a presença de um interlocutor e do público para acontecer, sendo o meio para sua execução um espaço físico ou virtual.

Pimenta e Almeida Júnior (2009, p. 92) definem mediação da informação como uma ação de interferência que pode e ocorre de maneira "*[...], direta ou indireta*; *consciente ou inconsciente*; *singular ou plural*; *individual ou coletiva*; *que propicia a apropriação de informação que satisfaça*, *plena ou parcialmente*, *uma necessidade informacional*".

Observa-se que a mediação sempre está envolvida na ação de informação, em algumas situações de forma mais ou menos explícita, inclusive nos processos de comunicação e divulgação científica; isto porque até mesmo o armazenamento e a representação (processos com mediação implícitos ou indiretos) envolvem em algum nível os interesses e as demandas do público (Pimenta; Almeida Júnior, 2009). Os autores decididamente distinguem mediação implícita e explícita; sendo que essa diferença reside no grau de interferência do mediador na ação. Quanto mais próximo do público e quanto mais interação exigir, mais explícito se torna este processo.

Sobre a comunicação científica via canais eletrônicos, Targino (2000) chama-nos a atenção que ela se configura como sendo de caráter semiformal por ter características formais e informais. Em seu aspecto informal, possibilita o contato entre os pesquisadores, além de favorecer "*[...] a troca rápida de informações e o feedback imediato ao desenvolvimento das pesquisas*". Quanto ao aspecto formal, "*favorece a divulgação do conhecimento produzido para um público amplo, em tempo menor do que a impressa*" (Oliveira; Noronha, 2005, p. 82).

Nessa direção, podemos categorizar nosso objeto de análise, a série *Diálogos de Pesquisa*, como um tipo de comunicação científica semiformal, além de constituir uma ação de mediação direta, onde o público pode interferir e interagir colocando suas observações, comentários e questionamentos.

Procedimentos metodológicos

O presente trabalho descreve e analisa os eventos desenvolvidos pela Comissão Diálogos de Pesquisa (CDP) do PPGCI/UFF na organização e realização do evento *Diálogos de Pesquisa*. Busca-se relatar a proposta da atividade, sua dinâmica, o perfil do público alcançado e suas impressões sobre o evento. Trata-se de estudo descritivo, de abordagem quanti-qualitativa, do tipo relato de experiência.

Segundo Gil (2002), o relato de experiência expressa um tipo de produção de conhecimento que qualifica os estudos descritivos, pois ajuda-nos a descrever fenômenos a partir do estabelecimento de relações de ação, sendo uma técnica relevante para a compreensão de fenômenos sociais, suas limitações e possibilidades de intervenção.

Na elaboração desta pesquisa, foram consultados o Portal de Periódicos da CAPES e a base de dados do Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação (Enancib), com o objetivo de verificar a existência de experiências similares relatadas em publicações e trabalhos científicos apresentados no Brasil. A escolha se deu por abarcarem textos brasileiros e estrangeiros, dando acesso a bases de dados, livros e periódicos, e aos anais do principal evento brasileiro em âmbito nacional da área da CI. Não foram localizados, porém, trabalhos semelhantes.

Dessa forma, a experiência da CDP não se restringe à organização de eventos de comunicação científica em plataforma de mídia social, mas também à estruturação de mecanismos de avaliação e divulgação da própria atividade. Esse relato de experiência, portanto, é uma primeira iniciativa de compartilhar e aprimorar a metodologia adotada até então.

A plataforma *Youtube*, ao disponibilizar métricas sobre os canais e conteúdos publicados, utiliza o termo *espectador* para referir-se às pessoas que assistem os vídeos ou acompanham as transmissões ao vivo. Neste artigo, optamos por utilizar o termo *participante*, considerando que a ideia de "ser espectador" despotencializa a proposta de mediação direta, cujo princípio está na interação. Os dados coletados foram apresentados e analisados a fim de compreender perfil, recepção e interação (via *chat*) dos participantes. Foram coletados dados de duas fontes: 1) estatísticas fornecidas pelo *YouTube* (duração do evento, pico de participantes simultâneos e número de visualizações) de todas as edições da *Diálogos de Pesquisa*; e 2) respostas em formulário elaborado no *Google Forms* que começou a ser compartilhado a partir de abril de 2023 no *chat* do Youtube durante a transmissão como requisito para emissão de certificado de ouvinte dos participantes. Os dados de todos os respondentes foram utilizados na análise.

Neste artigo, as *lives* foram identificadas por letras e por suas respectivas temáticas (Quadro 2) e não pelo nome dos palestrantes. Tal estratégia foi usada para garantir a anonimização de algumas informações referentes ao dia do evento.

O formulário foi construído com objetivo de compreender o perfil do público alcançado, além de obter um *feedback* sobre qualidade da divulgação, da transmissão e da apresentação dos palestrantes. O formulário sofreu alterações para atender às necessidades identificadas pela Comissão de traçar de maneira mais precisa o perfil do público e a efetividade das estratégias de difusão e do formato do evento. Considerando a intenção de aperfeiçoar os processos de divulgação, a configuração atual conta com oito perguntas divididas em dois blocos: *Presença no Evento* e *Sobre o Evento*. No primeiro, o público informa o nome completo, o seu perfil e a instituição afiliada. No segundo bloco, são levantados dados sobre o evento. O público, assim, responde sobre como soube do evento, o que achou da divulgação do evento, o que achou da qualidade da apresentação, o que achou da duração do evento e se pretende assistir a uma nova edição desta série de palestras em breve.

O período de apresentações contemplados nessa pesquisa se dá no intervalo entre os anos de 2020 e 2024. O formato remoto passou a ser adotado em função da pandemia de Covid-19, mantendo-se após o fim do isolamento social.

Resultados e discussão

O Programa de Pós-Graduação *Stricto Sensu* em Ciência da Informação da Universidade Federal Fluminense (PPGCI/UFF), foi criado em 2009. A área de concentração do Programa são as Dimensões Contemporâneas da Informação e do Conhecimento, que se desdobra em duas linhas de pesquisa: Informação, Cultura e Sociedade (enfoque social) e Fluxos e Mediações Sociotécnicas da Informação (enfoque em tecnologia e gestão). Conta com cursos de Mestrado Acadêmico (instituído em 2009) e doutorado (instituído em 2015) e oferece instâncias para pesquisas pós-doutorais (UFF, 2021).

Com vista a divulgar as pesquisas produzidas no âmbito do Programa, foi criada em 2020 a série de eventos *Diálogos de Pesquisa*. Em 28 de maio de 2020, em função da pandemia, o PPGCI criou um canal no *YouTube* e designou uma comissão para dar suporte à atividade: a CDP, composta por docentes e discentes.

As atividades da Comissão contemplam o suporte ao longo de todo o desenvolvimento das edições do evento, através das seguintes atribuições: 1) Captação de palestrantes; 2) Mediação e convite via *e-mail*; 3) Solicitação de artes para divulgação do evento nas redes sociais, listas de transmissão e

coordenações dos cursos de Arquivologia e Biblioteconomia e Documentação da instituição; 4) Criação da sala de transmissão através do *software* StreamYard e vinculação ao canal do programa no *YouTube*; 5) Elaboração de plataforma para gestão das inscrições e emissão de certificados via *Even3*; 6) Acompanhamento na realização dos eventos, através de suporte técnico e; 7) Envio dos certificados aos participantes e ouvintes.

Atualmente, a CDP conta com doze membros, sendo dez discentes e duas docentes. A dinâmica adotada funciona desta forma desde 2022.

O canal do PPGCI/UFF no *YouTube*, atualmente, conta com 58 vídeos, 1,19 mil inscritos e 19.757 visualizações (*Dados coletados em 16 de setembro de 2024*). O conteúdo do canal inclui vídeos que contemplam principalmente as seguintes categorias: 1) *Lives* apresentadas no decorrer da pandemia sobre saúde mental, cidadania e gestão de dados; 2) Gravações de seminários e encontros acadêmicos; 4) Aulas inaugurais; e 5) Edições da *Diálogos de Pesquisa* (objeto deste relato).

No Quadro 2, a seguir, foram descritas as apresentações organizadas pela Comissão entre agosto de 2020 e setembro de 2024. Ressalta-se que as oito primeiras *lives* descritas no quadro não tiveram dados coletados via *Google Forms*, por isso não foram incluídas nas análises sobre a afiliação e o perfil dos participantes. No entanto, achamos fundamental mencionar todas as palestras no Quadro 2, já que fazem parte da história deste evento.

Apresentação/*live*	Mês e ano	Temática	Tipo de evento
A	08/2020	Arquivos Pessoais	Seminário de encerramento de estágio pós-doutoral
B	10/2020	Censura	Seminário de encerramento de estágio pós-doutoral
C	07/2021	Comunicação Científica	Seminário de encerramento de estágio pós-doutoral
D	08/2021	Acesso à Informação	Encontro de pesquisas doutorais
E	12/2021	Acesso à Informação	Encontro de pesquisas doutorais
F	01/2022	Preservação Digital	Seminário de encerramento de estágio pós-doutoral
G	04/2022	Organização do Conhecimento	Seminário de encerramento de estágio pós-doutoral
H	05/2022	Ontologia	Seminário de encerramento de estágio pós-doutoral
I	04/2023	Ética do Arquivista	Encontro de pesquisas doutorais
J	07/2023	Acessibilidade em Arquivos	Encontro de pesquisas doutorais
K	09/2023	Gestão de Documentos	Seminário de encerramento de estágio pós-doutoral
L	11/2023	Neodocumentação	Encontro de pesquisas doutorais
M	12/2023	Acervos Bibliográficos e Memória	Seminário de encerramento de estágio pós-doutoral
N	12/2023	Organização do Conhecimento	Encontro de pesquisas doutorais
O	03/2024	Organização do Conhecimento	Encontro de pesquisas doutorais
P	04/2024	Organização da Informação	Encontro de pesquisas doutorais
Q	05/2024	Organização da Informação	Encontro de pesquisas doutorais
R	06/2024	Ciência Aberta	Encontro de pesquisas doutorais
S	07/2024	Arquivos Pessoais	Encontro de pesquisas doutorais
T	09/2024	Letramento Político	Relato de experiência

Quadro 2. Palestras em formato eletrônico organizadas pela CDP (agosto de 2020 a setembro de 2024)

Dessa maneira, ocorreram em 2020 duas edições, em 2021 e 2022, três, em 2023, seis; em 2024, até o mês de setembro, foram realizadas seis edições. Posto este panorama dos eventos, apresentamos, a seguir, os resultados da pesquisa, discutindo e analisando a dinâmica da *Diálogos de Pesquisa*, as estratégias de mediação mobilizadas pela CDP e o perfil e a recepção (interação e avaliação) do público alcançado nas transmissões.

Os eventos realizados pela CDP contemplam três categorias: 1) **Seminário de encerramento de estágio pós-doutoral**: eventos compostos pelo doutor que está realizando a apresentação, seu orientador e uma banca de avaliadores/debatedores que discutem o relatório apresentado; 2) **Encontro de pesquisas doutorais**: eventos de menor duração onde um recém-doutor, egresso do Programa, apresenta os resultados da sua pesquisa sob mediação de docente do PPGCI, preferencialmente o orientador; 3) **Relatos de experiência**: eventos em que discentes do PPGCI relatam práticas e/ou procedimentos de investigação, como as experiências em trabalho de campo e atividades de pesquisa, dentro e fora do país. Com base nessa categorização, observou-se a seguinte proporção no período relatado: 55% (onze apresentações) compõem os Encontros doutorais, 40% (oito apresentações) contemplam os Seminários de encerramento de estágio pós-doutoral e 5% (uma apresentação) se trata de relato de experiência de discente (Quadro 2).

A forma de mediação das *lives* organizadas pela CDP, com base em Pimenta e Almeida Júnior (2009), envolve uma estratégia de dupla mediação: direta (explícita) feita pelo mediador convidado (na maioria dos casos, o orientador da pesquisa apresentada), que situa o público sobre quem é o palestrante da edição e sua trajetória acadêmica, e indireta (implícita) realizada por meio de perguntas feita no *chat* disponibilizado, que são coletadas pela equipe de transmissão e repassadas ao mediador do evento, que as leva ao palestrante e interage com o mesmo.

Tal como em outros modelos de *lives*, os eventos promovidos pela CDP reúnem simultaneamente duas pessoas que interagem com o público: um mediador e um pesquisador, cabendo ao último a função de responder às perguntas dos espectadores A mediação tende a ser informal, aproximando-se mais de uma conversa. Como ressaltam Freitas e Rocha (2021, p. 6) "*[...] essa informalidade pode ser vista de maneira positiva, contribuindo para introduzir no imaginário coletivo a imagem do cientista como pessoa comum, ao contrário da imagem estereotipada que ainda hoje predomina*".

Em relação ao **total de visualizações** das transmissões salvas após os eventos apresentados, até outubro de 2024, observou-se uma média de 322 visualizações por vídeo. No entanto, é necessário mencionar a ocorrência de um ponto fora da curva, uma vez que a *live* Q alcançou 953 visualizações até a realização desta pesquisa.

Muitos fatores podem influenciar na popularidade de um vídeo: como a originalidade do conteúdo e a habilidade narrativa, tanto do mediador como do palestrante. Freitas e Rocha (2021) destacam, entre outros: linguagem compreensível ao público, popularidade do pesquisador, capacidade explicativa e descontração. No caso das lives organizadas pela CDP, é possível identificar a presença de pelo menos uma dessas características nos 20 vídeos disponíveis no canal do PPGCI/UFF no *YouTube*.

A **média de participantes simultâneos**, calculada com base nos dados coletados nas estatísticas fornecidas pelo *YouTube*, foi de 28,9 participantes (mínimo de 2 e máximo de 79 simultâneos). Foram excluídas deste levantamento a *live* Q, por ter sido considerada um dado *outlier*, diferenciando-se drasticamente dos demais, e a apresentação M, por ter sido realizada em auditório, não contando, portanto, com participantes simultâneos.

Em relação à **afiliação dos participantes**, observou-se presença majoritária de público interno, constituído por discentes das graduações de Arquivologia e de Biblioteconomia e da pós-graduação, além de docentes, técnicos e pesquisadores (71%). No entanto, cabe destacar a

presença de um público externo, de outras universidades (20%) e instituições de custódia (arquivos, bibliotecas e museus) (3%) (Gráfico 1).

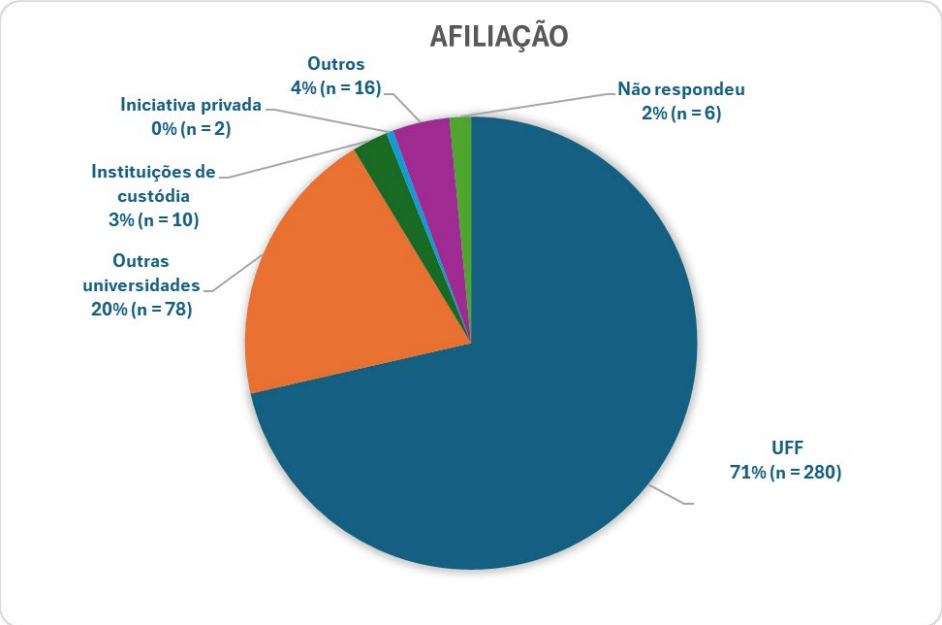


Gráfico 1. Instituição de afiliação.

A Tabela 1 descreve o **perfil dos participantes** alcançados em cada evento. Constatou-se que os discentes de graduação representam maioria em todas as edições, exceto na *live* K, que contou com uma presença expressiva de arquivistas. Somando os totais parciais contabiliza-se, nas 11 edições analisadas, 382 participantes ao vivo.

Apresentação/ *Live*	**Perfil dos participantes**						
	Discente Arquivologia	**Discente Biblioteconomia**	**Discente Pós-Graduação**	**Arquivista**	**Bibliotecário**	**Pesquisador**	**Outro**
I	17	5	6	0	0	0	1
J	16	9	5	0	0	0	5
K	12	3	7	14	1	0	1
L	9	11	7	0	4	0	2
N	12	5	2	0	5	0	3
O	9	31	10	0	0	0	3
P	4	10	9	0	0	0	1
Q	4	13	12	2	4	3	5
R	7	10	6	2	4	1	6
S	13	5	6	6	2	0	5
T	6	13	7	4	2	1	7
Totais/parciais	**109**	**115**	**64**	**28**	**22**	**5**	**39**

Tabela 1. Perfil dos participantes (2023-2024)

Sobre o **processo de divulgação**, incluindo os canais de comunicação usados pela Comissão (redes sociais, *e-mail*, convites pessoais, etc.) para atingir o público, as respostas se dividiram na seguinte proporção: 37% dos ouvintes declararam que a divulgação é excelente, 33% muito boa, 25% boa, 4% regular e 1% satisfatória ou ruim.

Com relação à **divulgação** (Gráfico 2), os meios mais expressivos foram Instagram (27%), WhatsApp (20%) e amigos e parentes (20%). O restante está distribuído entre e-mails e outras redes sociais.

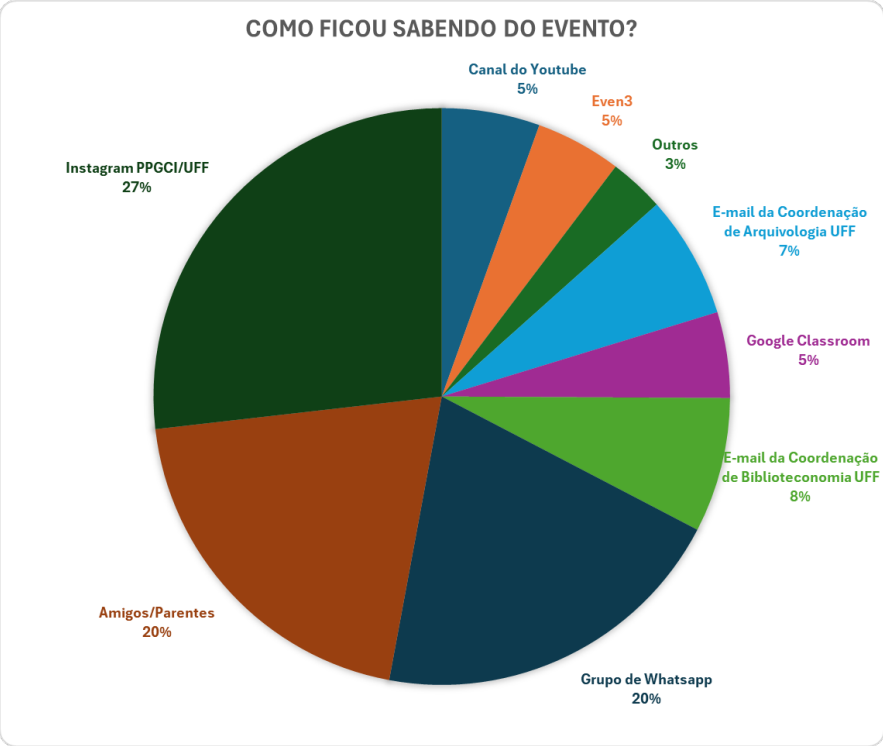


Gráfico 2. Como ficou sabendo do evento?

No que diz respeito a **assistir a uma nova edição do evento**, os participantes responderam, majoritariamente, que sim, assistiriam a outras edições da DP (97%). Outras respostas foram talvez (3%) e não (1 respondente).

Quanto às **interações do público** via *chat* do *YouTube*, foi obtida uma média de 3,3 perguntas por *live* (interação mínima de 0 e máxima de 10 perguntas). Duas *lives* não tiveram perguntas, visto que o(a) mediador(a) não abriu para questões, logo, não foram contabilizadas no cálculo da média de perguntas no *chat*. Quanto ao conteúdo dessas interações, geralmente envolvem questões metodológicas e conceituais.

Considerações finais

O acesso à comunicação científica, para além de bases indexadoras, catálogos *online* e bibliotecas digitais, vem sendo facilitado pelo desenvolvimento de novas tecnologias Web, como as plataformas de mídia social, como o *YouTube*, o que vem proporcionando a realização de eventos integralmente na modalidade digital, ou no formato híbrido, prática que se consolidou durante o período da pandemia de Covid-19.

No entanto, observa-se a prevalência desses formatos em muitos eventos acadêmicos, mesmo após o período pandêmico, o que revela uma nova cultura nos processos de comunicação científica do tipo informal no contexto brasileiro, sendo a *Diálogos de Pesquisa* um exemplo disso.

A mídia social *YouTube*, que já vinha tendo seu potencial explorado como ferramenta de divulgação científica, despontou como uma importante plataforma para mediação de informação científica. Atualmente, segundo Bueno (2011, p. 7), é possível vislumbrar iniciativas que facilitam a parceria entre comunicação e divulgação científica. A comunicação informal via canais eletrônicos, como o *YouTube*, tem potencial para facilitar a aproximação entre academia e sociedade, isto é, favorece a

transferência de conhecimento de forma mais interdependente. Apesar de percebermos que o público alcançado no caso em estudo é majoritariamente composto por acadêmicos e profissionais das áreas de Arquivologia e Biblioteconomia, a iniciativa de alimentar um canal oficial de programa de pós-graduação em uma plataforma tão popular quanto o *YouTube*, algo que outros programas de CI também vêm fazendo, é bastante válida. Além de serem recuperáveis, as *lives* viabilizam a troca de experiência entre os pares e podem alcançar um público leigo interessado.

Esta série de palestras tem como propósito apresentar as teses e projetos de pós-doutorado produzidos no PPGCI. Para além da banca e dos pesquisadores e estudantes interessados pela temática (público especializado), busca-se a difusão para um público mais amplo. Assim, um dos principais objetivos para a realização do projeto era o de aproximar a pós-graduação da graduação, considerando que a UFF abriga tanto um curso de Arquivologia, quanto um de Biblioteconomia. Os dados coletados demonstram que o objetivo está sendo alcançado, pois os estudantes desses cursos, juntos, constituem cerca de 60% do público levantado (Tabela 1). Esses dados levam-nos a aferir que o evento, apesar de se tratar de iniciativa de comunicação científica, também funciona como canal de divulgação científica, contribuindo para o letramento científico ao colocar alunos e alunas de graduação em contato com temas, objetos e metodologias aplicadas à área dos estudos da informação desenvolvidos no âmbito da pós-graduação. Constitui-se, além disso, como uma ação de mediação direta, aberta ao diálogo, questionamentos e reflexões entre palestrantes, mediadores e participantes.

Buscar-se-á, agora, a expansão do evento para fora dos muros da própria universidade. Por enquanto, apenas 1 a cada 4 ouvintes da *Diálogos de Pesquisa* é de outra universidade ou instituição. Para tanto, os dados referentes aos canais de divulgação do evento se mostram importantes indicadores para orientar uma autoavaliação da atividade. Assim, explorar diferentes maneiras de atingir um público mais amplo, estimular a participação e o diálogo entre pesquisadores e público, e difundir o conhecimento produzido na universidade fazem parte do compromisso da comissão responsável pelos eventos. Ademais, por se tratar de uma iniciativa de comunicação científica em canais informais, por meio de plataforma *online* privada, também desponta como preocupação formas de preservar não apenas as apresentações dos palestrantes, mas a interação do público e as trocas entre palestrantes, mediadores e ouvintes.

Este trabalho integra o conjunto de atividades de difusão e avanço do campo da Ciência da Informação desenvolvidas pela Comissão Diálogos de Pesquisa do PPGCI/UFF. Como identificado na análise da literatura existente sobre o tema e na pesquisa sobre outros relatos de experiência similares, essa pesquisa busca não só compartilhar as atividades e os resultados obtidos até então, mas também aperfeiçoar processos internos de avaliação e difusão do conhecimento produzido pela experiência prática da CDP e por seus membros. Desse modo, iniciativas como a de registrar as atividades realizadas, os dados levantados e estruturados até agora se mostram importantes para a preservação e divulgação das experiências vivenciadas, contribuindo para a difusão do conhecimento produzido por e para estudantes, pesquisadores e todos aqueles interessados nas diversas temáticas e objetos estudados em um programa de pós-graduação em Ciência da Informação.

Sobre os autores

Camilla Castro de Almeida é doutoranda e mestre em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal Fluminense (PPGCI-UFF). Graduada em Biblioteconomia e Documentação pela UFF. E-mail: camillaalmeida@id.uff.br

Juliana Maia Mendes é doutoranda e mestre em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal Fluminense (PPGCI-UFF). Graduada

em Arquivologia pela UFF e História pela Universidade Federal do Rio Grande do Norte (UFRN). E-mail: jmaia@id.uff.br

Vinicius Ribeiro Soares dos Santos é doutorando e mestre em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal Fluminense (PPGCI-UFF). Graduado em Biblioteconomia e Documentação pela UFF. Professor substituto no Departamento de Biblioteconomia da Universidade Federal do Estado do Rio de Janeiro (Unirio). E-mail: viniciusrsds@id.uff.br

Elisabete Gonçalves de Souza é professora do Departamento de Ciência da Informação e do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal Fluminense (UFF). Doutora em História e Filosofia da Educação Brasileira pela Universidade Estadual de Campinas (Unicamp), mestre em Educação pela UFF e graduada em Biblioteconomia e Documentação e em História pela UFF. E-mail: elisabetegs@id.uff.br

Lucia Maria Velloso de Oliveira é professora do Departamento de Ciência da Informação e do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal Fluminense (UFF). Doutora em História Social pela Universidade de São Paulo (USP), mestre em Ciência da Informação pelo Instituto Brasileiro de Informação em Ciência e Tecnologia e Universidade Federal Fluminense, graduada em História pela Universidade Federal do Rio de Janeiro (UFRJ) e em Arquivologia pela Universidade Federal do Estado do Rio de Janeiro (Unirio). E-mail: luciemarie@gmail.com

Referências

Brito, R.G. & Valls, V.M. (2015). Novas formas de aprendizagem e a mediação da informação: competências necessárias aos bibliotecários. *Revista Brasileira de Educação em Ciência da Informação*, 2(1), 3-28.

Bueno, W.C. (2010). Comunicação científica e divulgação científica: aproximações e rupturas conceituais. *Informação & Informação*, 15, 1-12.

Costa, J.L.O. (2009). *Padrões de comunicação em diferentes comunidades científicas*. Tese de Mestrado em Ciência da Informação. Braga: Universidade de Minho.

Cunha, R. B. (2017). Alfabetização científica ou letramento científico?: interesses envolvidos nas interpretações da noção de *scientific literacy*. *Revista Brasileira de Educação*, 22(68), 166-186.

Feitosa, L.T. (2016). Complexas mediações: transdisciplinaridade e incertezas nas recepções informacionais. *Informação em Pauta*, 1, 98-117.

Freitas, T.P. & Rocha, M.B. (2021). Lives de divulgação científica durante a pandemia: uma descrição do Instagram do Observatório Nacional. Encontro Nacional de Pesquisa em Educação em Ciências, 13, 2021. Anais... São Paulo: ABRAPEC.

Gil, A.C. (2022). Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas.

Oliveira, É.B.P.M. & Noronha, D.P. (2005). A comunicação científica e o meio digital. Informação & Sociedade, 15(1).

Pimenta, R.M. & Almeida Júnior, O.F. (2009). Mediação da informação e múltiplas linguagens. Tendências da Pesquisa Brasileira em Ciência da Informação, 2(1), 89-103.

Targino, M.G. (2000). Comunicação científica: uma revisão de seus elementos básicos. Informação & Sociedade, 10(2).

UFF [Universidade Federal Fluminense] (2020). Programa de Pós-graduação em Ciência da Informação. Diálogos de Pesquisa. Niterói-RJ.

UFF [Universidade Federal Fluminense] (2021). Resolução CEPEx/UFF Nº 214, de 14 de julho de 2021. Estabelece a alteração do Regimento Interno do Programa de Pós-Graduação em Ciência da Informação - Níveis mestrado e doutorado. Niterói-RJ.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Mapping GLAMs: creating a national dataset of GLAMs to develop a categorical climate-change risk assessment scale

Edward Benoit, III, Jill Trepanier, Jennifer Vanos, Emily Fisher, Kaitlyn Bailey, Annie Waddell, Haley Moore, Mandy Hatman, Mary Sidwell, Symonne Russell, Virginia Seger, Paige Boutte, Amanda Latta, Zoe Mohammad, Kyriel Felton, Erin Deliman, Breanna Benson-Pearce, Wendy Johnson, Allyson Russell, Baillie Pretzer, Christopher Reeder, Melissa McConnell, Lisa Dahlke, Kaitlynn Melear, Lillian Bodi, Savannah T. Lyle, Zach Lannes, Gwen L. Wells, Benjamin A. Teincuff, Jason M. Straight, Tiffany Rockwell, Shane T. Manthei, Jennifer L. Benner, Jane Fiegel, Amanda Lima, Elizabeth Rininger, Caroline Melinger, Deborah Metz-Andrews, Meryl Roepke, Karen Isaac, and Mallory Collins

DOI: https://doi.org/10.47989/ir30iConf47185

Abstract

Introduction. The PROTECCT-GLAM project aims to assess and address climate risks for U.S. galleries, libraries, archives, and museums (GLAMs). The project team began with creating a national dataset of GLAMs.

Method. The project team used existing datasets that required different auditing and manipulation techniques to align its data. Following the data normalization, the master file included 77,960 entries.

Analysis and Results. The project team used ArcGIS Pro to analyze a dataset of galleries, libraries, archives, and museums (GLAMs) for risk from sea level rise, finding 16,877 GLAMs within 30 kilometres of the U.S. coast, with California, New York, Maryland, Florida, and New Jersey having the highest numbers. They also assessed GLAMs per capita by state using 2020 census data, reporting results per 10,000 people.

Conclusion. The dataset creation was successful and is being utilized to create a five-point scale based on the average of five climate variables for each GLAM.

Introduction

For decades, climate scientists have attempted to warn society about potential significant threats related to a changing climate as an attempt to call for action. Unfortunately, many did not take the danger's immediacy seriously, and the risks have now moved from potential to present. According to the 2021 Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report, widespread and rapid changes in the atmosphere, ocean, and biosphere have already occurred. The predominant evidence of these changes can be noticed in extreme behavior of heatwaves, heavy precipitation, droughts, and tropical cyclones, which will all continue to increase in severity and frequency into the future (IPCC, 2023). The nine U.S. regions considered by the Fourth National Climate Assessment all have increased threats from these events into the future, regardless of which future climate scenario is considered (Black & Brown, 2020).

More than ever, cultural heritage institutions must assess their unique climate change-related threats to their collections and missions as part of their disaster and emergency management plans. According to the 2019 heritage health information (HHI) survey, 58% of institutions do not have a disaster plan, with 75% lacking both a plan and staff training to implement a plan (IMLS, 2019). Furthermore, the HHI report indicated an increased attribution of damage between 2017 and 2019 due to water or moisture (56%), with natural disasters accounting for 10% (IMLS, 2019).

The PROTECCT-GLAM project attempts to address these challenges by developing a national categorical climate change risk assessment scale for galleries, libraries, archives, and museums (GLAMs). To meet this objective, the project team must first address the lack of a comprehensive geo-referenced dataset for all U.S. GLAMs. The following paper discusses this dataset's development and initial GIS analysis—specifically, identifying the number of GLAMs located within 5, 15, and 30 kilometers of a coastline and the number of GLAMs per capita in each state.

Literature review

Previous GIS analysis of GLAMs remains limited and explores the factors at play in the location of cultural institutions, specifically public libraries, rather than the potential impacts of climate change. Wan, Chen, & Zhang (2017) sought to develop a statistical model to aid in the placement of new libraries. In their model, they account for traffic, environmental factors, a library's constraint value, and the capacity limitation of the library. Similarly, Shorabeh et al. (2020) developed three major categories to consider when deciding where to build a new public library: centrality (population and accessibility), compatibility (air and noise pollution and comfort and security), and environmental (elevation, slope, and fault lines). Their findings indicate that optimal locations for public libraries are among areas with high population density, with ample access to transportation (nearby roadways and public transit) and security centers (police and fire stations, health resources, etc.).

Further, their results also showed that population and demographics were more critical in location decisions than accessibility, meaning that building libraries in areas with higher rates of illiterate population and student density is more beneficial than building libraries near roads or public transportation. Findings from a study investigating the relationship between accessibility and the utilization rate of public libraries support the conclusions of Shorabeh et al. (Zhao & Hong, 2023). In contrast with results from former studies, there was no evidence of a causal relationship between accessibility and utilization; in comparison, the findings showed that human factors, like motivation and customs and habits, play a more important role in public library utilization among their sample.

The location of cultural institutions has important implications for their surrounding areas and those seeking to use their resources. In particular, the existence and construction of new museums have been shown to be a catalyst for growth in the area in which they are located. For example, in a study of the proximity of museums to the city centers of Barcelona, Paris, and Turin, (Paül i Agustí,

2014) found that while most museums are located within 2 km of the city center, those outside this radius have more significant impacts on their surrounding communities than those located within the radius and can be used by community leaders to spur social and economic change within their reach. These findings illustrate cultural institutions' role in their communities and their potential growth.

Despite these hopeful findings, however, cultural institutions often follow suit of other significant social and economic resources, meaning that areas with scarce access to these resources generally lack access to cultural institutions as well (Short et al., 2020). In their investigation of informal learning institutions (ILIs) locations across the United States, Short et al. found several '*ILI deserts*' worthy of attention. In particular, their findings show large areas among the Great Plains, the Southeast, and the Northwest with a low relative density of ILIs compared to their population. Further, these '*deserts*' often encompass areas with high poverty rates and low population density and disproportionately affect American Indian and Alaskan Native communities, with 0.95% of these communities lacking access to ILIs compared to the general population's 0.10%. This problem is not exclusive to these areas, however. In a study of Washington DC, Baltimore, Chicago, and Phoenix and their respective spatial accessibility of public libraries, Cheng et al. (2021) found that, with the exception of Phoenix, socially disadvantaged populations, including non-adults, minorities, and impoverished people, have to travel further distances to access public libraries. Combined, these studies indicate that demographics facing other societal and economic hardships across the country are less likely to have easy access to cultural institutions.

Methodology

To meet the PROTECCT-GLAM project's overall objectives, the project team needed to create a dataset that included the geographic, physical location of all GLAMs in the U.S. The project team began constructing the dataset by identifying the following existing external data files:

- 2018 Museum Data Files from IMLS Data Catalog (IMLS, 2018)
- 2020 Public Libraries Survey (PLS) from IMLS Data Catalog (IMLS, 2020)
- 2019 Archives RepoData at GitHub (Goldman, Tansey, & Ray, 2019)
- 2012 Academic Library Survey (ALS) Public Use Data File from the National Center for Education Statistics (NCES, 2012)
- 2023 Department of Defense Morale, Welfare and Recreation (MWR) Libraries Directory (US DOD, 2023)

Each external data file required different auditing and manipulation techniques to align its data to the PROTECCT-GLAM requirements. The following section summarizes the steps taken for each external data file.

Museum data files

The museum data files were last updated in 2018 and include three separate data files. The project team merged the individual files into a single data file with 30,177 entries. (Note: The project team discovered several duplicate entries later in the process, accounting for the disparity between the original number of entries and the final number in the data file.) The Museum Data File Documentation and User's Guide noted that the data file included 56 informational columns (Frehill & Pelczar, 2018). The original data included information based on IRS nonprofit data, IMLS administrative data, private foundation grants, professional organizations, and other public sources.

The project team removed all data columns except COMMONNAME, DISCIPL, GSTREET, GCITY, GSTATE, GZIP, GZIP5, LONGITUDE, and LATITUDE and assigned all entries a unique identification

with the prefix MU. The team then identified entries that did not include formal location addresses (e.g., those with P.O. Boxes or incomplete addresses) and entries without geo-references. This process marked 5,290 entries that required additional auditing and verification.

Public libraries survey data files

The project team utilized the most recently available Public Libraries Survey from fiscal year 2020. The source included three files. Of the three files, only the Public Library Outlet Data File included 17,509 public library service outlets. The Public Libraries Survey Fiscal Year 2020 Data File Documentation and User's Guide noted that the outlet data file included 41 informational columns (Pelczar et al., 2022). The original data included information based on state- and territory-based reporting units.

The project team removed all data columns except STABR, LIBID, LIBNAME, ADDRESS, CITY, ZIP, CNTY, LONGITUD, and LATITUDE and assigned all entries a unique identification with the prefix PL. The team then identified entries that did not include formal location addresses (e.g., those with P.O. Boxes or incomplete addresses) and entries without geo-references. This process marked 142 entries that required additional auditing and verification.

Repository data (RepoData) for United States archives

The project team downloaded the most recent RepoData JSON file on August 30, 2022, from GitHub. The source comprises a single data file with 25,771 unique entries and 24 informational columns. The original data included information from 138 archival organizations collected between 2017 and 2019 by Ben Goldman, Eira Tansey, and Whitney Ray (2019).

Since RepoData is an active data file and to provide the RepoData team with updated entries, the project team retained all of the original data columns. It assigned all entries a unique identification with the prefix ARC. The team then identified entries that did not include formal location addresses (e.g., those with P.O. Boxes or incomplete addresses) and entries without geo-references. This process marked 8,132 entries that required additional auditing and verification.

Academic libraries data files

The project team utilized the most recently available Public Libraries Survey, which was published in 2014. The team exported the Microsoft Access information into an Excel data file with 4,261 unique entries. As noted in the *documentation for the academic libraries survey* (ALS) *Public Use Data File: Fiscal Year 2012*, the data file included 180 informational columns (Phan, Hardesty, & Hug, 2014). The original data included information collected by the National Center for Education Statistics (NCES) with assistance from the American Library Association's Office of Research and Statistics.

The project team removed all data columns except UNITID, INSTNM, ADDR, CITY, STABBR, and ZIP, assigning all entries a unique identification with the prefix AL. The team then identified entries that did not include formal location addresses (e.g., those with P.O. Boxes or incomplete addresses) and entries without geo-references. This process marked all 4,261 entries as having formal addresses without geo-reference.

MWR library data

The department of defense morale, welfare, and recreation (MWR) Libraries data were extracted from the DOD MWR website directory. The directory included 12 separate fields; the resulting data file included 196 unique entries. The project team removed the 67 libraries outside the United States, resulting in 129 entries. Additionally, the team removed all data columns except: BASE; LIBRARY; ADDRESS; CITY; STATE; and POSTALCODE and assigned all entries a unique identification with the prefix MWR. None of the directory entries included geo-references.

Data verification and updating

The project team extracted the 17,954 entries requiring additional verification or updating from each of the data files:

- Museum: 5,290
- Public Libraries: 142
- Archives: 8,132
- Academic Libraries: 4,261 (only required georeferenced)
- MWR Libraries: 129 (only required georeferenced)

Members of the project team reviewed and corrected each entry, followed by a secondary quality review by two team members. Limited entries whose physical location could not be verified were removed from the data file. Following the verification process, the individual data files were recompiled with the updated entries.

Data file compilation

The team identified the required data elements for the final PROTECCT-GLAM data file based on the available information shared by most GLAM data files and normalized the individual data files accordingly. The final data file included the following elements: a unique identifier, the original data source assigned identifier (if available), the original data source, the repository name, street address, city, state or territory, zip code, county or parish, repository type (each entry is assigned a primary and secondary type), latitude and longitude of the repository, the user id of who last updated the entry, and date of last revision.

Combined data file

Following the data normalization, the project team compiled all the data files into a single master file, resulting in 77,960 entries. The team conducted a final review to remove 4,602 duplicate entries. During the review process, the research team retained duplicate entries listed within both museum and archival repositories with the interpretation that the archival entry refers to the museum's archive. The final PROTECCT-GLAM data file includes 73,358 unique entries.

Analysis and results

After its creation for initial visualization and spatial analysis, the project team imported the dataset into ArcGIS Pro. The initial analysis focused on GLAMs most at risk for potential sea level rise by identifying all GLAMs located within 5, 15, and 30 kilometers of the U.S. coast. As noted in Table 1, a combined total of 16,877 GLAMs are within 30 km of the coast. The states with the highest number of GLAMs within 30 km are California (2,672), New York (2,566), Maryland (1,304), Florida (1,290), and New Jersey (1,159).

Repository Type	5 km	15 km	30 km
Archives	2458	3702	4740
Libraries	2054	3555	4625
Galleries & Museums	3915	5886	7512
Total	8427	13143	16877

Table 1. Number of GLAMs located within 30 km of a U.S. coast

Secondary analysis focused on the number of GLAMs per capita in each state using the 2020 U.S. census data. For this analysis, the per capita is reported at the per 10,000 level (see Table 2).

State	Archives	Libraries	Galleries & Museums	All GLAMs
ME	8.43	2.12	3.77	14.32
VT	3.19	3.23	4.20	10.62
ND	3.41	1.40	3.20	8.01
NH	1.44	1.86	2.42	5.72
IA	1.56	1.97	1.84	5.38
DC	1.74	0.65	2.42	4.81
NE	1.60	1.53	1.65	4.78
AL	2.95	0.73	0.83	4.51
WY	0.21	1.51	2.77	4.49
SD	0.19	1.85	2.36	4.40
KS	1.02	1.47	1.62	4.12
AR	2.25	0.90	0.95	4.09
MT	0.49	1.29	2.24	4.02
AK	0.33	1.45	2.07	3.85
NY	2.08	0.67	1.00	3.76
RI	1.43	0.77	1.49	3.69
WV	1.15	1.13	1.28	3.56
PA	1.60	0.65	1.15	3.41
MS	1.62	0.92	0.78	3.32
MA	0.94	0.83	1.37	3.13
NM	0.99	0.81	1.27	3.06
WI	0.78	0.92	1.34	3.03
OK	1.23	0.70	1.08	3.01
MN	1.02	0.82	1.10	2.93
MO	0.86	0.79	1.12	2.76
CT	0.29	0.79	1.38	2.45
LA	0.65	0.85	0.76	2.27
OH	0.46	0.77	1.04	2.27
ID	0.30	0.91	1.04	2.25
NC	0.91	0.52	0.70	2.14
MI	0.45	0.74	0.92	2.11
OR	0.25	0.66	1.18	2.09
IL	0.35	0.74	0.90	1.99
HI	0.34	0.50	1.10	1.94
SC	0.63	0.56	0.69	1.88
VA	0.30	0.56	0.98	1.84
DE	0.15	0.47	1.21	1.84
IN	0.12	0.77	0.91	1.81
GA	0.63	0.49	0.58	1.71
NJ	0.37	0.55	0.75	1.67
CO	0.09	0.59	0.99	1.67
KY	0.08	0.62	0.93	1.63
MD	0.23	0.41	0.93	1.57
WA	0.11	0.57	0.82	1.51
AZ	0.44	0.41	0.56	1.42
TN	0.16	0.54	0.70	1.40
TX	0.31	0.38	0.56	1.25
UT	0.26	0.48	0.46	1.21
CA	0.08	0.40	0.61	1.09
FL	0.12	0.34	0.48	0.94
NV	0.02	0.34	0.50	0.86

Total	**0.68**	**0.64**	**0.90**	**2.21**

Table 2. Number of GLAMs per 10,000 people in each state

Conclusion and future directions

The PROTECCT-GLAM project team successfully created a national dataset of GLAMs and conducted a preliminary analysis using ArcGIS Pro. Additional analysis will incorporate data on five variables, including hurricane wind occurrence & storm surge inundation, three feet of sea level rise, inland flooding occurrence, temperature exceeding 23-27^{0}C occurrence, and dew point exceeding 9.3^{0}C occurrence. The resulting data will be compiled into a five-point scale based on the average of the five variables for each GLAM.

Limitations

The information within the data file was based on external data files, and all associated limitations apply to the combined data file. Although the project team reviewed and updated data entries, the fluid nature of repositories inherently leads to out-of-date information. More importantly, community based GLAMs and those from underrepresented communities are not adequately included in the initial data file.

The PROTECCT-GLAM project team will address these limitations by openly distributing the data file and requesting that GLAM community members submit missing institutions or revisions to included institutions via the project website.

Acknowledgements

This study was funded by the Institute of Museum and Library Services and with the support of the Louisiana State University School of Information Studies.

About the authors

Edward Benoit, III is Associate Professor and Interim Director of the School of Information Studies at Louisiana State University. He received his Ph.D. from the University of Wisconsin-Milwaukee, and his research interests are in community and participatory archives, climate change impacts on libraries, archives, and museums, and archival pedagogy. He can be contacted at ebenoit@lsu.edu

Jill Trepanier is Professor and Chair of the Department of Geography and Anthropology at Louisiana State University. She received her Ph.D. from Florida State University and specializes in hurricane climatology and the estimate of risk using statistical methods. She can be contacted at jtrepa3@lsu.edu

Jennifer Vanos is Associate Professor in the School of Sustainability and College of Global Futures at Arizona State University. She received her Ph.D. from the University of Guelph, Canada. Her research addresses adaptation to extreme heat in a changing climate and the impacts of extreme heat and air pollution on human health. She can be contacted at jenni.vanos@asu.edu

Emily Fisher is a master's student in the Department of Geography and Anthropology at Louisiana State University. She earned her B.S. from the University of West Florida. Her research focuses on understanding impacts from extreme weather events on vulnerable communities. She can be contacted at efish24@lsu.edu

Kaitlyn Bailey is doctoral student in the College of Art & Design at Louisiana State University. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Her research focuses on GIS applications for historical analysis in cultural heritage spaces. She can be contacted at kbail54@lsu.edu

Annie Waddell earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Haley Moore is a Records Analyst with Access Sciences. She earned her M.L.I.S. from Louisiana State University.

Mandy Hatman is a graduate student in the Records and Information Management program at Louisiana State University. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University. She can be contacted at mcome41@lsu.edu

Mary Sidwell is an intern at the State Archives of North Carolina. She earned her M.L.I.S. from Louisiana State University.

Symonne Russell is a Case Coordinator for the Office of Guardian ad Litem in Colorado Springs. She earned her M.L.I.S. from Louisiana State University.

Virginia Seger is an Educational Assistant at Barlett City Schools. She earned her M.L.I.S. from Louisiana State University.

Paige Boutte is a Reference Librarian at the Louisiana Secretary of State. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Amanda Latta is an Assistant Librarian and Archivist for Swisher Community Library, Iowa, and an Assistant for the Iowa Conservation and Preservation Consortium. She earned her M.L.I.S. from Louisiana State University.

Zoe Mohammad is an undergraduate student in Coastal Environmental Science at Louisiana State University. She may be contacted at zmuham6@lsu.edu

Kyriel Felton earned her M.L.I.S. and Graduate Certificate in Records & Information Management from Louisiana State University.

Erin Deliman is a Collections Management Intern at the Nation Afro-American Museum and Cultural Center, Wilberforce, Ohio. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Breanna Benson-Pearce is an Animal Registrar for the Utah Hogle Zoo. She earned her M.L.I.S., Graduate Certificate in Archival Studies, and Graduate Certificate in Records and Information Management from Louisiana State University.

Wendy Johnson is Director of Library Services and e-Learning Coordinator for River Parishes Community College in Louisiana. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Allyson Russell earned her M.L.I.S. from Louisiana State University.

Baillie Pretzer is an Administrative Support Assistant at the Berkeley County Library. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Christopher Reeder is a graduate student in the M.L.I.S. and Graduate Certificate in Records & Information Management programs at Louisiana State University. He may be contacted at creede8@lsu.edu

Melissa McConnell is a graduate student in the M.L.I.S. and Graduate Certificate in Archival Studies programs at Louisiana State University. She may be contacted at mmcco74@lsu.edu

Lisa Dahlke is an Archivist at the Amistad Research Center in New Orleans. She earned her Graduate Certificate in Records & Information Management from Louisiana State University.

Kaitlynn Melear is an undergraduate student in Coastal Environmental Science at Louisiana State University. She may be contacted at kmelea1@lsu.edu

Lillian Bodi is an undergraduate student in Biological Engineering at Louisiana State University. She may be contacted at lbodi1@lsu.edu

Savannah T. Lyle earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Zach Lannes is Engineering Librarian at Stanford University. He earned his M.L.I.S. from Louisiana State University.

Gwen L. Wells is Archivist at the Jones Memorial Library in Lynchburg, Virginia. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Benjamin A. Teincuff is a Systems Analyst for the American Association for the Advancement of Science. He earned his M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Jason M. Straight is a Research Analyst at Morgan, Lewis and Bockius in Boston, MA. He earned his M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Tiffany Rockwell is a Senior Information Technology Support Analyst at LSU Health Sciences Center, New Orleans. She earned her M.L.I.S. from Louisiana State University.

Shane T. Manthei is a Librarian at the Free Library of Philadelphia. He earned his M.L.I.S. from Louisiana State University.

Jennifer L. Benner is a Librarian at the Tampa-Hillsborough County Public Library. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Jane Fiegel is a Metadata Librarian at Tulane University. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Amanda Lima is a Library Assistant at the Bodleian Libraries, University of Oxford. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Elizabeth Rininger earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Caroline Melinger is a Collections Manager for AWG Art Advisory in New York City. She earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Deborah Metz-Andrews earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Meryl Roepke is a Research Assistant at the Orcas Island Historical Society, Eastsound, Washington. She earned her M.L.I.S. from Louisiana State University.

Karen Isaac earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

Mallory Collins earned her M.L.I.S. and Graduate Certificate in Archival Studies from Louisiana State University.

References

Black, A.W., & Brown, V. M. (2020). *Inventory and probabilistic function of natural hazards in the United States: July 19, 2018 - September 18, 2018*. National Renewable Energy Laboratory. https://doi.org/10.2172/1595574

Cheng, W., Wu, J., Moen, W., & Hong, L. (2021). Assessing the spatial accessibility and spatial equity of public libraries' physical locations. *Library and Information Science Research*, 43(2), 1-9. https://doi.org/10.1016/j.lisr.2021.101089

Frehill, L. M. and Pelczar, M. (2018). *Data file documentation: Museum data files, FY 2018 release*. Institute of Museum and Library Services.

Goldman, B., Tansey, E., & Ray, W. (2019). *Repository data (RepoData) for United States archives*. https://github.com/RepoData/RepoData

Institute of Museum and Library Services (IMLS) (2018). *Museum data files*. https://www.imls.gov/research-evaluation/data-collection/museum-data-files

Institute of Museum and Library Services (IMLS) (2019). *Protecting America's collections: Results from the heritage health information survey*. IMLS.

Institute of Museum and Library Services (IMLS) (2020). *Public libraries survey*. https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey

Intergovernmental Panel on Climate Change (IPCC) (2023). *Climate change 2023: Synthesis report* (H. Lee & J. Romero, Eds.). IPCC. https://doi.org/10.59327/IPCC/AR6-9789291691647

National Center for Education Statistics (NCES) (2012). *Library statistics program*. https://nces.ed.gov/surveys/libraries/aca_data.asp

Paül i Agustí, D. (2013). Differences in the location of urban museums and their impact on urban areas. *International Journal of Cultural Policy*, 20(4), 471–495. https://doi.org/10.1080/10286632.2013.850498

Pelczar, M., Soffronoff, J., Nielsen, E., Li, J., & Mabile, S. (2022). *Data file documentation: Public libraries in the United States fiscal year 2020*. Institute of Museum and Library Services.

Phan, T., Hardesty, L., & Hug, J. (2014). *Documentation for the Academic Libraries Survey (ALS) public-use data file: Fiscal year 2012* (NCES 2014-040). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Shorabeh, S.M., Varnaseri, A., Firozjaeri, M.K., Nickravesh, F., & Samany, N.N. (2020). Spatial modeling of areas suitable for public libraries construction by integration of GIS and multi-attribute decision making: Case study Tehran, Iran. *Library and Information Science Research*, 42(2), 1-12. https://doi.org/10.1016/j.lisr.2020.101017

Short, R. A., Struminger, R., Zarestky, J., Pippin, J., Wong, M., Vilen, L., & Lawing, A. M. (2020). Spatial inequalities leave micropolitan areas and Indigenous populations underserved by informal STEM learning institutions. *Science advances*, 6(41), eabb3819. https://doi.org/10.1126/sciadv.abb3819

United States Department of Defense (2023). *Morale, welfare, and recreation libraries directory*. https://www.dodmwrlibraries.org/locations

Wan, B., Chen, Q., & Zhang, J.-r. (2017). Analysis on the location of community library based on nationwide reading. *16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*, 88-92. https://doi.org/10.1109/DCABES.2017.26

Zhao, X. & Hong, K. (2023). Basic analysis of the correlation between the accessibility and utilization activation of public libraries in Seoul: focusing on location and subway factors. *Buildings*, 13(3), 1-16. https://doi.org/10.3390/buildings13030600

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Promoting AI literacy through U.S. academic libraries: an analysis of LibGuides from ARL and Oberlin group libraries using the EDUCAUSE AI literacy framework

Ko Chun Ru and Rong Tang

DOI: https://doi.org/10.47989/ir30iConf47182

Abstract

Introduction. As the integration of artificial intelligence (AI) rapidly advances, academic libraries are increasingly pivotal in supporting AI literacy among students and faculty.

Method. Through content analysis, the present study examines 70 newly developed generative AI LibGuides from academic libraries affiliated with the association of research libraries (ARL) and the Oberlin group, using the EDUCAUSE AI literacy framework.

Analysis. Through a detailed examination, the present research reorganizes and improves the EDUCAUSE AI literacy framework, proposing a more comprehensive version tailored to higher education needs. The adapted framework fills the gaps in the original model and offers a nuanced approach to AI literacy, reflecting the unique challenges faced by academic libraries.

Results. The findings reveal that most LibGuides emphasize foundational AI tools and responsible use, with less focus on advanced technical competencies related to AI creation. Significant differences were observed between ARL and Oberlin Group LibGuides, with ARL offering more comprehensive coverage. To address these differences, consistent training and knowledge sharing initiatives are recommended to ensure a common standard of AI literacy support across academic libraries.

Conclusion. This study provides insights into the role of libraries in promoting generative AI literacy and identifies areas for future strategic partnerships and improvement.

Introduction

As the integration of artificial intelligence (AI) rapidly advances, academic libraries are increasingly pivotal in supporting AI literacy among students and faculty (Mairn, 2024). This study provides a novel examination of how U.S. academic libraries, particularly those affiliated with the ARL and Oberlin Group, address generative AI through their library guides (hereafter referred to as LibGuides). Employing content analysis guided by the EDUCAUSE AI Literacy Framework, this study focuses on 70 recently created LibGuides dedicated to generative AI.

The originality of this study lies in the adaptation of the EDUCAUSE AI Literacy Framework, developed for Barnard College to guide AI literacy in educational settings (Hibbert et. al., 2024), within the context of academic library guides. This framework, which comprises four levels: *understanding* AI, *use and apply* AI, *analyse and evaluate* AI, and *create* AI, was reorganized to align more closely with the specific characteristics and objectives of LibGuides.

The present study holds strong implications for advancing AI literacy in higher education. By reorganizing and enhancing the EDUCAUSE AI literacy framework, it proposes a more comprehensive framework tailored specifically to the needs of higher education institutions. This adaptation addresses gaps identified in the original framework, offering a nuanced approach to AI literacy that reflects the unique challenges and opportunities faced by academic libraries.

Additionally, the study highlights on the evolving role of librarians in the higher education setting and their collaboration with faculty, AI experts, and other campus stakeholders in promoting AI literacy. Understanding how librarians and campus stakeholders may share responsibilities in disseminating AI knowledge provides valuable insights into how these roles can work together more effectively. This may pave the way for future research on optimizing collaboration between librarians' and other campus members' roles to address the broader impact of AI.

This study also identifies limitations in librarians' current approach, such as the often-limited technical depth in their guides. These findings highlight the need for further research into librarians' competencies and the future training required to meet the growing demands of AI literacy education. Addressing these limitations is crucial for enhancing the effectiveness of library resources and services while empowering librarians in their key role in advancing AI literacy.

Literature review

Generative artificial intelligence (AI) creates new images, text, and software by learning from existing data without directly replicating it. Although AI development has been ongoing for years, the launch of ChatGPT in November 2022 captured global attention due to its ease of use through natural language interaction. While generative AI offers considerable convenience, it also raises critical concerns, including issues of transparency, accuracy, bias, intellectual property disputes, cybersecurity risks, and sustainability challenges (Gartner, 2023). The rapid proliferation of AI tools for writing, coding, translation, information retrieval, illustration, and presentation creation also poses challenges in higher education. Not all campus stakeholders are familiar with or have access to these tools (Henke, 2024), potentially exacerbating disparities in learning opportunities. Furthermore, while it is recognized for enhancing productivity, concerns have surfaced regarding its effectiveness in creative and intellectually demanding fields, as well as its potential to facilitate plagiarism, which has heightened academic integrity concerns (Mogavi et al., 2024). According to EDUCAUSE's '*AI blueprint report*', a survey of 910 higher education professionals conducted in late 2023, 78% of respondents believe generative AI will impact academic integrity and the design of assignments, assessments, and curricula. Additionally, 73% reported that their institutions have already implemented related strategies, and over half have started offering education and guidance on the subject (Robert, 2024).

Libraries have long played a vital role in supporting teaching and research in higher education, serving as key partners in the adoption of technology and information literacy education. In '*Developing a library strategic response to artificial intelligence*,' Cox (2023) conducted a survey of 111 libraries in the UK and identified the perceived impacts of AI on various library functions. These include using AI to enhance metadata, improve discovery and retrieval processes, and implementing AI-driven tools, such as chatbots for user services. The paper outlines three key strategies for libraries:

> *using library AI capabilities to model responsible and explainable applications of descriptive AI, leveraging Librarians data competencies to enhance organizational AI capability, and promoting AI literacy to enhance organizational and societal AI capabilities.*

AI literacy is emerging as a critical area of focus for libraries around the world, driven by the increasing integration of AI technologies into information services and academic environments. Many information literacy instructors have begun exploring ways to transition into teaching AI literacy (Hervieux & Wheatley, 2024). Additionally, numerous librarians have designed AI literacy courses based on the core principles of information literacy and shared their approaches (ACRL, 2024).

An increasing number of academic librarians are developing library guides tailored to specific subjects or courses, assisting users with database searches, and recommending relevant resources. These guides serve as valuable tools for teaching information literacy and can, at times, have a greater impact than traditional classroom instruction (Duffy et al., 2021). Ko and Chiu (2024), in their analysis of 28 library guides using the ACRL framework, also found that librarians effectively applied the core concepts of information literacy to the introduction of AI. Currently, these guides present a timely opportunity to help build, promote, and advance AI literacy. By building on their foundation in information literacy, librarians can provide essential support for higher education institutions as they navigate broader AI policy development and respond to the growing demands of the AI era.

AI literacy framework

The concept of AI literacy was first introduced in 2015, and scholars from various fields have continued to explore its definition. At its core, AI literacy centres on equipping individuals with the ability to critically assess, engage with, and make use of AI technologies, even if they do not have the skills to develop AI models themselves (Chan, 2024). According to a review by Lintner (2024), although many AI literacy scales have been developed, their reliability and validity still require further verification. These scales share common elements, particularly in their emphasis on a basic technical understanding of AI, awareness of its societal implications, and a focus on AI ethics.

Meanwhile, Talagala (2021) introduced the concept of context of AI application, suggesting that individuals from different fields and with varying levels of prior knowledge will require different aspects of AI literacy. Pinski and Benlian (2024) provide an overarching conceptual framework for AI literacy, synthesising findings from 68 academic papers. Their work critically examines the specificity of AI literacy to different user groups and distinguishes it from other technology literacies. The conceptual framework is structured around the categories of '*Proficiency Dimensions*' and '*Subject Areas*', which encompasses the content areas relevant to AI literacy: AI models; data for AI; AI interfaces; AI tools; humans, organizations, and society; and cross-area. The cross area includes various literacy skills, including ethical literacy, critical literacy, meta-literacy, and future literacy.

Ng et al. (2021) explored the concept of AI literacy and provided a foundation for defining, teaching, and evaluating it in educational contexts. Through reviewing 18 peer-reviewed articles published

between 2016 and 2021, the authors identified four key aspects of AI literacy: '*know and understand* AI', '*apply* AI,' '*evaluate and create* AI,' and '*AI ethics*'. The first three of these components were aligned with Bloom's taxonomy. A few years later, Ng et al. (2024) developed an AI literacy questionnaire (AILQ) and proposed a model based on the ABCE (affective, behavioural, cognitive and ethical) to evaluate students' literacy development. The authors' framework lays the foundation for other frameworks, including the EDUCASE AI literacy framework, offering educators information on how to foster AI literacy in students.

In June 2024, EDUCAUSE published its AI literacy framework, offering a structured approach to understanding AI literacy in educational contexts. This framework is designed to guide institutions in developing AI education programs and is intended to meet individuals where they are scaffolding upon their current level of AI literacy. These levels are not strictly sequential; individuals can engage at any level based on their prior knowledge and goals (Hibbert et. al., 2024). The framework breaks down AI literacy into four levels:

Levels	Details
Level 1. Understand AI	Focuses on building a basic understanding of AI, including its definitions, history, and core technologies.
Level 2. Use and apply AI	Focuses on applying AI in practical contexts, such as using AI-powered software, services, or platforms.
Level 3. Analyse and evaluate AI	Focuses on developing critical thinking skills to analyse AI systems, evaluate their performance, and assess their societal and ethical impacts.
Level 4. Create AI	Focuses on advanced skills for creating AI models, such as developing custom machine learning models or AI-driven applications.

Table 1. EDUCAUSE AI literacy framework

Research questions

The following are the research questions our study is pursuing:

RQ1. For librarians developing AI literacy LibGuides, how could the EDUCAUSE AI Literacy framework be modified to be optimally suitable for libraries' user needs?

RQ2. To what extent are the Levels and Core Competencies (CCs) and Key Concepts (KCs) in the EDUCAUSE AI literacy framework addressed by the 70 ARL and Oberlin institutions' LibGuides? What levels and specific CCs/KCs are most commonly covered? What levels and specific CCs/KCs are the least frequently covered?

RQ3. What differences exist in terms of the coverage of EDUCAUSE AI literacy framework between Oberlin and ARL libraries and among the four levels?

Method

This study employs content analysis, guided by the EDUCAUSE AI literacy framework, to examine LibGuides created by academic libraries in the USA affiliated with the ARL and the Oberlin Group. The focus is exclusively on newly created guides specifically addressing generative AI. Guides in which librarians simply added information to existing content or those that are simply compilations of resources without explanatory text were excluded from the analysis, as they were deemed ineffective in conveying AI literacy. Furthermore, to avoid potential disciplinary biases, guides that introduced generative AI from a subject-specific angle were excluded.

By reviewing 114 ARL members in the United States and 83 members of the Oberlin group, the researchers identified 48 ARL libraries and 22 Oberlin libraries that created new LibGuides for AI, meeting the criteria for analysis. The content of these 70 guides was downloaded by the researchers, both having backgrounds in library and information science, between June 24 and June 30. The subsequent analysis was conducted between July and August 2024.

Among the 70 libraries analysed, several produced multiple guides addressing different aspects of AI, while others consolidated their information in a single guide. For this study, each library was treated as a unit of analysis to capture the overall message conveyed through its guides (details of the study samples are provided in Appendix 1).

The EDUCAUSE AI literacy framework consists of four levels, each linked to specific core competencies, key concepts, and reflection questions. However, the framework does not explain how these elements were developed. To address this limitation, the researchers reorganized the framework, elevating some key concepts and reflection questions to core competencies. Similar or overlapping concepts were consolidated, and multidimensional core competencies were split into sub-codes. Those not elevated or consolidated were categorized as sub-codes under their respective core competencies. Additionally, the researchers incorporated concepts from the LibGuides that were missing or underrepresented in the original framework, reflecting the LibGuides' unique perspective, and enriching the coding schema for a more comprehensive analysis. Tables 2, 3, 4, and 5 show the core competency, sub-code, description, and source for Level 1, Level 2, Level 3, and Level 4, respectively (examples of coding details are provided in Appendix 2).

To validate the coding scheme, a pre-test was conducted on a sample of 10% of the guides (seven guides) to ensure intercoder reliability, yielding an agreement ratio at 83.97%. Throughout this process, MAXQDA and Excel were used as coding tools, and consensus on newly emerging codes was achieved through discussion.

Core Competency	Sub-Code	Description	Source
L1 CC1 Be able to define AI-related terms	L1 CC1-1	Define the term '*Artificial Intelligence*'	EC-L1-KC1
	L1 CC1-2	Define the term '*Generative Artificial Intelligence*'	LG
	L1 CC1-3	Define the term '*Machine Learning*'	EC-L1-KC1
	L1 CC1-4	Define the term '*(Large) Language model*'	
	L1 CC1-5	Define the term '*Artificial Neural Network*'	
	L1 CC1-6	Define the term '*Diffusion Model*'	
	L1 CC1-7	Define the terms '*Artificial narrow intelligence*,' '*Artificial general intelligence*,' '*Artificial super intelligence*,' '*Reactive machines*,' '*Theory of mind*,' '*Self-awareness*'	EC-L1-KC2
	L1 CC1-8	Define technical frameworks related to AI (open source vs. closed models, APIs, etc.).	EC-L1-KC4
	L1 CC1-9	Recognize AI tool categories	EC-L1-KC3, RQ4
	L1 CC1-10	Define the term '*Prompt*'	LG
	L1 CC1-11	Other	

L1 CC2 Recognize the benefits and limitations of AI tools	L1 CC2-1	Identify technologies utilized by AI tools	EC-L1-RQ2
	L1 CC2-2	Understand the designed purpose of AI tools	EC-L1-RQ3
	L1 CC2-3	Assess what AI tools are particularly useful and what they are not	EC-L1-RQ4
	L1 CC2-4	Compare the benefits and limitations of AI tools to databases/search engines	LG
L1 CC3 Identify and explain the differences between various types of AI, as defined by their capabilities and computational mechanisms	L1 CC3-1	Identify the differences between various types of AI	EC-L1-CC
	L1 CC3-2	Explain the differences based on AI capabilities	
	L1 CC3-3	Describe the differences based on the AI computational mechanisms	
	L1 CC3-4	Explain the differences based on AI fees/access limits	LG
	L1 CC3-5	Use alternative methods to explain differences in AI types	
L1 CCP Additional competencies for understanding AI	L1 CCP-1	Track AI news or events	LG
	L1 CCP-2	Find AI resources	
	L1 CCP-3	Recognize AI working groups, associations, or institutions	
	L1 CCP-4	Understand the history of AI	
	L1 CCP-5	Recognize developments in AI-augmented tools	
	L1 CCP-6	Recognize AI literacy issues	

- EC-CC refers to the original core competencies outlined in the EDUCAUSE framework.
- EC-KC refers to the original key concepts outlined in the EDUCAUSE framework.
- EC-RQ refers to the original research questions outlined in the EDUCAUSE framework.
- LG refers to the librarian perspectives displayed in the LibGuides.
- Note. L1 CCP is the new competency identified in LibGuides.

Table 2. Level 1 '*identify* AI' coding scheme

Core Competency	Sub-Code	Description	Source
L2 CC1 Successfully utilize generative AI tools for desired responses	L2 CC1-1	Used solely to code AI usage methods beyond prompting.	EC-L2-CC
L2 CC2 Experiment with	L2 CC2-1	Apply '*Prompt Engineering*' or '*Context Windows*', '*Zero-Shot Prompting*', '*Few-Shot Prompting*'	EC-L2-KC1
	L2 CC2-2	Prompting technique- Add specificity	EC-L2-

prompting techniques to improve AI-generated output	L2 CC2-3	Prompting technique - Use context and details	KC2
	L2 CC2-4	Prompting technique - Ask the model to consider pros and cons	
	L2 CC2-5	Prompting technique - Ask the model to evaluate alternative positions	
	L2 CC2-6	Prompting techniques to reduce bias and hallucinations	EC-L2-KC1, RQ3
	L2 CC2-7	Prompting technique - Utilize iterative prompting	LG
	L2 CC2-8	Prompting technique -Task segmentation	
	L2 CC2-9	Prompting technique - Limit prompt word count or length	
	L2 CC2-10	Prompting technique - Use polite phrasing	
	L2 CC2-11	Prompting technique - Provide examples in prompts	
	L2 CC2-12	Other Prompting techniques	
	L2 CC2-13	Apply Prompting framework	
	L2 CC2-14	Apply Prompting template	
L2 CC3 Review AI-generated content	L2 CC3-1	Review content for hallucinations	EC-L2-KC1
	L2 CC3-2	Review content for fake citations	LG
	L2 CC3-3	Review content for potential reasoning errors	EC-L2-KC4
	L2 CC3-4	Review content for how prompts may affect outcomes	LG
	L2 CC3-5	Review content for potential bias	
	L2 CC3-6	Review content for reliability	
	L2 CC3-7	Review content for accuracy	
	L2 CC3-8	Review content for currency issue	
	L2 CC3-9	Strategies for reviewing AI-generated content	
	L2 CC3-10	Methods for detecting AI-generated content	
L2 CC4 Considerations for information fed into prompting tools	L2 CC4-1	Privacy consideration	EC-L2-KC3
	L2 CC4-2	Confidentiality consideration	
	L2 CC4-3	Copyright consideration	
L2 CCP Proper AI usage	L2 CCP-1	Apply institutional AI policy	LG
	L2 CCP-2	Disclose AI use in classroom work	
	L2 CCP-3	Apply publisher AI policy	

disclosure	L2 CCP-4	Disclose AI use in scholarly writing	
	L2 CCP-5	Cite AI-generated content in the APA, MLA, and Chicago style	
	L2 CCP-6	Keep a record of AI prompts for future verification	
	L2 CCP-7	Other ways to disclose AI usage	

- The same EC-CC, EC-KC, EC-RQ and LG notes as Table 2.
- Note. L2 CC4 is described as KC3 in the EDUCAUSE framework but is not explicitly listed as a Core Competency in the original framework. L2 CCP is the new competency identified in LibGuides.

Table 3. Level 2 '*use and apply AI*' coding scheme

Core Competency	Sub-Code	Description	Source
L3 CC1 Examine AI in a broader context, bringing in knowledge from one's discipline or interests	L3 CC1-1	Examine AI in a broader context	EC-L2-CC
	L3 CC1-2	Bring knowledge from specific disciplines or interests	
	L3 CC1-3	Framework for evaluating AI tools	LG
L3 CC2 Critique AI tools and offer arguments in support of or against their creation, use, and application	L3 CC2-1	Critical Perspective -Reliability	LG
	L3 CC2-2	Critical Perspective -Closed model or black box structure	
	L3 CC2-3	Critical Perspective -Overreliance and its effects on learning	
	L3 CC2-4	Evaluation of the designer's intentions versus actual usage of AI tools	
	L3 CC2-5	Strategies to prevent creators' work being used as AI training data	
	L3 CC2-6	Guidelines for developing AI policies in the classroom	
	L3 CC2-7	Other	
L3 CC3 Analyse ethical considerations in AI development and deployment	L3 CC3-1	Ethical Considerations -Environmental sustainability	EC-L3-KC1
	L3 CC3-2	Ethical Considerations -Labor	
	L3 CC3-3	Ethical Considerations -Privacy	
	L3 CC3-4	Ethical Considerations -Copyright	
	L3 CC3-5	Ethical Considerations -Biases (race, gender, class and other)	
	L3 CC3-6	Ethical Considerations -Misinformation	
	L3 CC3-7	Ethical Considerations -Academic integrity	LG
	L3 CC3-8	Ethical Considerations -Authorship concerns	
	L3 CC3-9	Ethical Considerations -Information Privilege/Digital Divide	
	L3 CC3-10	Other	
• The same EC-CC, EC-KC, EC-RQ and LG notes as Table 2.			

Table 4. Level 3 '*analyse and evaluate* AI' coding scheme

Core Competency	Sub-Code	Description	Source
L4 CC1 Synthesize learning to conceptualize or create new ideas, technologies, or structures that relate to AI	L4 CC1-1	Conceive of novel uses for AI	EC-L4-CC
	L4 CC1-2	Build software leveraging AI	
	L4 CC1-3	Propose theories about AI	
L4 CC2 Human-AI collaboration	L4 CC2-1	Examine how your ideas differ from AI-generated content	EC-L4-RQ1
	L4 CC2-2	Identify unique AI features that contribute to creativity	
L4 CCP Access resources to support AI creation	L4 CCP-1	Access resources for creating new AI-driven ideas or technologies	LG

- The same EC-CC, EC-KC, EC-RQ and LG notes as Table 2.
- Note. L4 CC2 is described as RQ1,2 in the EDUCAUSE framework, but it is not explicitly listed as a Core Competency in the original framework. L4 CCP is the new competency identified in LibGuides.

Table 5. Level 4 '*create* AI' coding scheme

Findings and discussion

In this section, we present the findings of our study based on the research questions outlined earlier.

Audience and professional background shape differences between the EDUCAUSE framework and LibGuides

Analysing library guides through the lens of the EDUCAUSE framework reveals two distinct approaches to AI literacy. While the EDUCAUSE framework is primarily designed for classroom instruction, librarians focus on curating and providing access to a diverse array of resources. Among the 70 library guides analysed, only two adopted a module format similar to online courses, incorporating unit activities or reflective exercises. Furthermore, many guides extend beyond student audiences, offering support to educators by helping them integrate AI into their teaching practices and adapt to advancements in AI. However, when presenting AI use cases, guides more frequently highlight applications for learning rather than specific support for teaching (see Table 6).

AI Use Situations	%
To assist learning	64.3
To assist teaching	47.1
To assist research	45.7
To assist search	24.3
To assist writing	22.9
To assist language translation	15.7
Other use case/situation	15.7

Table 6. AI use cases in library guides and their frequency of occurrence

Library guides prioritize the accessibility and diversity of resources, as reflected in the introduction of new core competencies at Level 1. For example, '*additional competencies for understanding* AI' encompass topics such as the history of AI, tracking AI news, recognizing key AI organizations, and locating AI resources (see Table 2). This resource-oriented approach underscores librarians' distinctive role to fostering AI literacy and meeting the evolving needs of higher education.

Beyond resource accessibility, librarians have introduced a critical Core Competency to Level 2 (Use and Apply AI): '*the proper disclosure of AI usage*' (see Table 3). Drawing on their expertise in citation practices, librarians emphasize this competency to ensure compliance to regulations such as copyright laws, publisher policies, institutional guidelines, citation standards, and AI privacy policies. This focus on AI disclosure addresses the academic integrity challenges posed by AI, which were not explicitly covered in the original EDUCAUSE framework.

Leveraging their deep foundation in information literacy, librarians underscore the differences between AI tools and their appropriate use, aligning with the ACRL information literacy framework's concept of '*information creation as a process*'. However, the primary focus remains on tools for text generation and research support (see Table 7), while fewer guides address AI-related media literacy topics, leaving a notable gap for future development.

AI tools categories	%
LLM or Chatbot, e.g. : ChatGPT, Gemini, Claude	61.4
Assisting research AI tools	51.4
Image generation AI tools	40
Search AI tools	30
Other types of AI tool, for example, Alexa, Deep Blue, and predictive text	28.6
Coding AI tools	27.1
Writing AI tools	24.3
Video generation AI tools	24.3
AI content detection tools	24.3
Text generation AI tools	21.4
Audio generation AI tools	20
Document analysis AI tools	10
Automatic creations of slides AI tools	8.6
Language translation or correction AI tools	7.1
Creating designs and mock-ups AI tools	4.3
Automatic transcription AI tools	1.4

Table 7. Categories of AI tools mentioned in library guides and their frequency of use

Librarians also highlight critical perspectives on information use, reflecting the ACRL Information Literacy Framework's principles of '*authority is constructed and contextual*' and '*Information Has Value.*' In the Level 2 section on reviewing AI-generated content, some LibGuides extend beyond basic concerns such as fake citations and outdated information by incorporating verification methods like the SIFT framework. These guides further enrich Level 3 (analyse and evaluate AI) by exploring essential topics, including information privilege, the digital divide, overreliance on AI tools, and the alignment of AI designers' intentions with actual use.

While these contributions reflect the depth of librarians' expertise, certain limitations remain evident. Frameworks such as CLEAR, ROBOT, and SIFT are frequently repeated across various guides, indicating a strong culture of shared learning within the library community. However, this repetition may also signal challenges in swiftly integrating AI-related frameworks from other disciplines.

Moreover, despite introducing numerous critical perspectives, notable gaps exist. For instance, issues such as the potential impact of AI on reducing interpersonal interactions and diminishing individual creativity are underexplored (Stahl, 2024). Additionally, most guides focus on understanding current AI issues and cautioning users about associated risks, rather than empowering them to actively shape AI-related policies. This gap underscores a potential area for growth in Level 4 (create AI).

Coverage of the EDUCAUSE framework

As shown in Table 8, the levels most frequently covered within the EDUCAUSE AI Literacy framework are Level 1 (understand AI) and Level 3 (analyse and evaluate AI). The top three Core Competencies with the highest occurrence rates among the 70 library guides are: '*L1 CC2: recognize the benefits and limitations of AI tools*' (41.1%), '*L1 CCP: additional competencies for understanding AI*' (35.7%), and '*L3 CC3: analyse ethical considerations in the development and deployment of AI*' (35.3%). These reflect the librarian's focus on foundational AI knowledge and ethical considerations.

In contrast, the least covered levels are Level 4 (Create AI) and parts of Level 2 (Use and Apply AI). The core competencies with the lowest occurrence rates are '*L4 CC1: synthesize learning to conceptualize or create new ideas, technologies, or structures related to AI*' (1.1%), '*L4 CC2: human-AI collaboration*' (2%), and '*L2 CC1: successfully utilize generative AI tools for desired responses*' (4.2%). This suggests that librarians may have limited expertise in these areas or may not fully recognize their importance as integral components of AI literacy.

Core Competencies by Levels	%
L1 CC2 Recognize the benefits and limitations of AI tools	41.1
L1 CCP Additional competencies for understanding AI	35.7
L3 CC3 Analyse ethical considerations in the development and deployment of AI	35.3
L3 CC1 Examine AI in a broader context, bringing in knowledge from one's discipline or interests	35.1
L1 CC1 Be able to define the AI related terms	28.6
L2 CC3 Review AI-generated content	28
L2 CCP Proper AI usage disclosure	27.9
L2 CC4 Considerations for information fed into prompting tools	17.8
L1 CC3 Identify and explain the differences between various types of AI, as defined by their capabilities and computational mechanisms	15.2
L2 CC2 Experiment with prompting techniques to improve AI-generated output	14.9
L4 CCP Access resources to support AI creation	13.2
L3 CC2 Critique AI tools and offer arguments in support of or against their creation, use, and application	12.8
L2 CC1 Successfully utilize generative AI tools for desired responses	4.2
L4 CC2 Human-AI collaboration	2
L4 CC1 Synthesize learning to conceptualize or create new ideas, technologies, or structures that relate to AI.	1.1

Table 8. Frequency of core competencies by levels in library guides

A detailed analysis of the sub-codes provides additional insight. Among the 70 library guides, 13 key concepts were mentioned in more than half of them. As shown in Table 9, the most frequently addressed were: '*L1 CC1-9: recognize AI tool categories*' (85.7%), '*L1 CCP-2: find AI resources*' (77.1%),

and '*L2 CCP-5: cite AI-generated content in APA, MLA, and Chicago styles*' (74.3%). In contrast, 19 key concepts appeared in fewer than 10% of the guides. The concepts least frequently introduced in the library guides are: '*L4 CC1: synthesize learning to conceptualize or create new ideas, technologies, or structures that relate to AI*' (1.1%), '*L4 CC2: human-AI collaboration*' (2%), and '*L2 CC1: successfully utilize generative AI tools for desired responses*' (4.2%) (See Table 10).

Key concepts mentioned in more than half of the guides	%
L1 CC1-9 Recognize AI tool categories	85.7
L1 CCP-2 Find AI resources	77.1
L2 CCP-5 Cite AI-generated content in APA, MLA, and Chicago style	74.3
L3 CC3-5 Ethical Considerations -Biases (race, gender, class and other)	65.7
L1 CC2-3 Assess what AI tools are particularly useful for and what they are not	62.9
L2 CC3-1 Review content for hallucinations	62.9
L1 CC1-2 Define the term "Generative Artificial Intelligence"	55.7
L2 CCP-2 Disclose AI use in classroom work	55.7
L3 CC3-7 Ethical Considerations -Academic Integrity	55.7
L2 CC3-7 Review content for accuracy	51.4
L2 CCP-1 Apply institutional AI policy	51.4
L3 CC3-4 Ethical Considerations -Copyright	51.4
L1 CCP-3 Recognize AI working groups, associations, or institutions	50

Table 9. Frequency of key concepts mentioned in more than half of the library guides

Key concepts mentioned in fewer than 10% of the guides	%
L1 CC1-7 Define the term '*Artificial narrow intelligence*,' '*Artificial general intelligence*,' '*Artificial super intelligence*,' '*Reactive machines*,' '*Theory of mind*,' or '*Self-awareness*'	8.6
L1 CC3-5 Use alternative methods to explain differences in AI types	8.6
L2 CC2-9 Prompting technique - Limit prompt word count or length	8.6
L3 CC2-5 Strategies to prevent creators' work being used as AI training data	8.6
L1 CCP-5 Recognize developments in AI-augmented tools	7.1
L1 CC1-8 Define technical frameworks related to AI (open-source vs. closed models, APIs, etc.).	5.7
L2 CC1 Successfully utilize generative AI tools for desired responses	5.7
L2 CC2-6 Prompting techniques to reduce bias and hallucinations	5.7
L3 CC2-4 Evaluation of designers' intentions versus actual usage of AI tools	5.7
L2 CC2-4 Prompting technique - Ask the model to consider pros and cons	2.9
L2 CC2-10 Prompting technique - Use polite phrasing	2.9
L4 CC2-1 Examine how your ideas differ from AI-generated content	2.9
L2 CC2-5 Prompting technique - Ask the model to evaluate alternative positions	1.4
L2 CC3-3 Review content for potential reasoning errors	1.4
L4 CC1-2 Build software leveraging AI	1.4
L1 CC1-6 Define the term "Diffusion Model"	0
L4 CC1-1 Conceive of novel uses for AI	0
L4 CC1-3 Propose theories about AI	0
L4 CC2-2 Identify unique AI features that contribute to creativity	0

Table 10. Frequency of key concepts mentioned in less than 10% of the library guides

Overall, the analysis reveals that LibGuides tend to focus on building a foundational understanding of AI tools and encouraging responsible use of AI, while dedicating less attention to the technical aspects and competencies in Level 4, which involve AI creation. These trends suggest that the guides prioritize helping users discover additional resources for continued learning and upholding academic integrity standards, rather than training them in advanced AI creation, programming, or operational skills (detailed coding concept distribution and applications across four AI literacy levels are provided in Appendix 3).

Statistical differences

The descriptive statistics of AI-related LibGuides' coverage for both Oberlin group (OBL) libraries and association of research libraries (ARL) were analysed based on the four levels of core competencies from the EDUCAUSE AI literacy framework (see Table 11).

Levels	Library Group	Mean	SD	Median	Min	Max
Level 1: Understand AI	OBL	0.248	0.213	0.205	0.000	0.773
	ARL	0.333	0.219	0.323	0.000	0.896
Level 2: Use and Apply AI	OBL	0.174	0.138	0.182	0.000	0.500
	ARL	0.216	0.177	0.157	0.000	0.688
Level 3: Analyse and Evaluate AI	OBL	0.209	0.172	0.159	0.000	0.636
	ARL	0.303	0.178	0.240	0.083	0.667
Level 4: Create AI	OBL	0.000	0.000	0.000	0.000	0.000
	ARL	0.013	0.019	0.000	0.000	0.042

Table 11. Library guides' coverage measured by EDUCASE AI literacy framework levels

For OBL, the mean coverage for Level 1 was 0.248 (SD = 0.213). For ARL, the mean coverage for Level 1 was higher at 0.333 (SD = 0.219). For OBL, the mean coverage for Level 2 was 0.174 (SD = 0.138). Whereas for ARL, the mean coverage for Level 2 was 0.216 (SD = 0.177).

For Level 3, analyse and evaluate AI, the mean coverage of OBL was 0.209 (SD = 0.172). In comparison, ARL libraries' mean coverage for Level 3 was higher at 0.303 (SD = 0.178). For Level 4, the numbers are different: while OBL libraries had no coverage for this level, with a mean of 0.000 (SD = 0.000), whereas ARL libraries had a low mean coverage at 0.013 (SD = 0.019).

Differences between OBL and ARL libraries

A paired samples t-test was conducted to compare the percentage coverage of EDUCAUSE AI literacy framework components between Oberlin and ARL libraries. The results showed that the mean percentage coverage for OBL (M = 0.192, SD = 0.175) was significantly lower than that of the ARL libraries (M = 0.258, SD = 0.199), with paired differences as ($t(72) = -4.861$, $p < .001$). The effect size ranged from moderate to large ($d = -0.569$), indicating a significant difference in the AI literacy coverage between the two library groups. The mean difference in coverage was -0.066 (SD = 0.117), suggesting that the ARL libraries' LibGuides generally covered more components of the AI literacy framework than the LibGuides developed by libraries of the Oberlin group.

Differences among AI literacy levels

The average proportion of coverage was obtained by averaging OBL and ARL's proportion of coverage of various core competencies and key concepts. The mean proportion of coverage was highest for Level 1 (M = 0.291, SD = 0.206), followed by Level 3 (M = 0.256, SD = 0.163), Level 2 (M = 0.195, SD = 0.149), and the lowest for Level 4 (M = 0.006, SD = 0.009).

A one-way ANOVA was conducted to examine the effect of different levels on the average proportion of LibGuides' coverage of the EDUCAUSE AI literacy framework competency components. One-way ANOVA revealed a statistically significant difference in the average proportion of LibGuides coverage across the different levels ($F(3, 69) = 4.436$, $p = .007$). The effect size was small to moderate, as indicated by eta-squared ($\eta^2 = 0.162$), suggesting that approximately 16.2% of the coverage variance can be attributed to the level differences.

Post hoc tests using Tukey's HSD indicated that the mean coverage for Level 4 was significantly lower than for Level 1 (mean difference = -0.284, $p = .006$) and Level 3 (mean difference = -0.250, $p = .019$). No other comparisons were statistically significant.

Conclusions

This study examines 70 newly developed generative AI LibGuides from libraries affiliated with ARL and Oberlin group, analysed through the lens of the EDUCAUSE AI literacy framework. Our findings align with Talagala's (2021) assertion that AI literacy must adapt to individuals' roles and task contexts. Similarly, those responsible for providing AI literacy adapt their focus based on their training, background, and tasks. LibGuides include content absent from the EDUCAUSE framework, such as proper disclosure of AI usage, highlighting academic integrity concerns. Furthermore, LibGuides emphasize Level 1 (understand AI) and Level 3 (analyse and evaluate AI), focusing on helping users critically evaluate AI tools' strengths and limitations.

However, many technical concepts featured in the EDUCAUSE framework, such as '*diffusion models*' and '*few-shot prompting*', were underrepresented in the LibGuides. Furthermore, prompt techniques and operational instructions for AI tools at Level 2 (use and apply AI) were less frequently addressed. The near absence of Level 4 (create AI) was particularly striking, suggesting that libraries focus more on resource discovery and ethical AI use, rather than advanced AI creation and programming.

These findings underscore the need for collaboration across higher education to foster comprehensive AI literacy, with various roles contributing their unique expertise. Hervieux and Wheatley's (2024) AI literacy framework, grounded in Bloom's taxonomy and shaped by a library science perspective, provides a concise alternative to the EDUCAUSE framework. By integrating their approach with the detailed competencies identified in this study, institutions could develop a more actionable and inclusive AI literacy framework that better addresses diverse user needs. Moreover, the detailed coding scheme from this study offers libraries a valuable tool for identifying underrepresented areas and fostering partnerships with other campus units to bridge these gaps effectively.

Our analysis also revealed that ARL libraries generally cover more components of the AI literacy framework than Oberlin group libraries, which may reflect differences in resources, institutional priorities, and librarian training. Addressing these disparities through more consistent training and knowledge-sharing initiatives could ensure a common standard of AI literacy support across academic libraries.

Lastly, this study excludes discipline-specific generative AI guides, which could offer valuable insights into discipline-based AI literacy needs. Friesen et al. (2023) examined publications by librarians across various disciplines and found that legal and health sciences librarians had the highest number of AI-related articles, with different disciplines focusing on distinct AI-related topics. Future research should explore whether different disciplines prioritize unique AI literacy competencies and how librarians in specific fields can better support those needs.

Acknowledgements

We used GPT-4.0 in part for translation and proof reading. The output was evaluated through authors' review and rewriting. The authors assume full responsibility for the content of this submission.

About the authors

Ko Chun Ru is a Doctoral student at the Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan. Her research interests include information literacy, and open education. She can be contacted at 81115003e@ntnu.edu.tw

Rong Tang is Professor Emerita at Simmons University's School of Library and Information Science. She received her Ph.D. from University of North Carolina at Chapel Hill. Her research interests are in research data management, open government data, UX research, and information behaviour. She can be contacted at rong.tang@simmons.edu

References

ACRL (2024). *ACRL Framework for Information Literacy Sandbox.* https://sandbox.acrl.org/resources-search?keys=Artificial+intelligence

Chan, C. K. Y., & Colloton, T. (2024). AI literacy. In Chan, C. K. Y., & Colloton, T. (Ed.), *Generative AI in Higher Education: The ChatGPT Effect* (pp. 24-43). Routledge. https://doi.org/10.4324/9781003459026

Cox, A. (2023). *Developing a Library Strategic Response to Artificial Intelligence* [White paper]. IFLA Artificial Intelligence SIG. https://www.ifla.org/g/ai/developing-a-library-strategic-response-to-artificial-intelligence/

Duffy, B., Maluski, K. & Levitan, G. (2021). Framing the Guides: Transforming LibGuides Creation Through Conceptual Integration with the ACRL Framework. In J. L. Finch (Ed.), *Envisioning the Framework: A Graphic Guide to Information Literacy* (pp. 23-41). Association of College and Research Libraries.

Friesen, E., Tanna, H. & Roy, A. (2023). Artificial Intelligence in Subject-Specific Library Work: Trends, Perspectives, and Opportunities. *Canadian Journal of Academic Librarianship*, 9, 1–26. https://doi.org/10.33137/cjal-rcbu.v9.39951

Gartner. (2023). *Gartner Experts Answer the Top Generative AI Questions for Your Enterprise*. https://www.gartner.com/en/topics/generative-ai

Henke, J. (2024). Navigating the AI era: university communication strategies and perspectives on generative AI tools. *Journal of Science Communication*, 23(03), A05. https://doi.org/10.22323/2.23030205

Hervieux, S. & Wheatley, A. (2024). Building an AI Literacy Framework: Perspectives from Instruction Librarians and Current Information Literacy Tools [White paper]. https://www.choice360.org/research/white-paper-building-an-ai-literacy-framework-perspectives-from-instruction-librarians-and-current-information-literacy-tools/

Hibbert, M., Altman, E., Shippen, T. & Wright, M. (2024). A Framework for AI Literacy. https://er.educause.edu/articles/2024/6/a-framework-for-ai-literacy

James, A. B. & Filgo, E. H. (2023). Where does ChatGPT fit into the Framework for Information Literacy? The possibilities and problems of AI in library instruction. *College & Research Libraries News*, 84(9), 334-341.

Ko, C. R. & Chiu, M. H. (2024). How Can Academic Librarians Support Generative AI Literacy: An Analysis of Library Guides Using the ACRL Information Literacy Framework. *Proceedings of the Association for Information Science and Technology*, 61(1), 977-979. https://doi.org/10.1002/pra2.1159

Lintner, T. (2024). A systematic review of AI literacy scales. *npj Science of Learning*, 9, 50 https://doi.org/10.1038/s41539-024-00264-4

Mogavi, R. H., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., Tlili, A., Bassanelli, S., Bucchiarone, A., Gujar, S., Nacke, L. E., & Hui, P. (2024). ChatGPT in education: A blessing or a

curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behaviour: Artificial Humans*, 2(1). https://doi.org/10.1016/j.chbah.2023.100027

Ng, D. T. K., Leung, J. K. L., Chiu, T. W. S., & Qiao, M. S. (2021). AI Literacy: Definition, Teaching, Evaluation and Ethical Issues. *Proceedings of the Association for Information Science and Technology*, 58(1) 504-509, https://doi.org/10.1002/pra2.487

Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2024). Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive, and ethical approach. *British Journal of Educational Technology*, 55, 1082–1104. https://doi.org/10.1111/bjet.13411

Pinski, M., & Benlian, A. (2024). AI literacy for users – A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behaviour: Artificial Humans*, 2(1), 100062. https://doi.org/10.1016/j.chbah.2024.100062

Robet, J. (2024). *EDUCAUSE AI Landscape Study*. EDUCAUSE. https://library.educause.edu/resources/2024/2/2024-educause-ai-landscape-study

Stahl, B.C. & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74, 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Talagala, N. (2021, April 4). The four Cs of AI literacy: Building the workforce of the future. Forbes. Retrieved 15 September 2024 from https://www.forbes.com/sites/nishatalagala/2021/04/04/the-four-cs-of-ai-literacy-building-the-workforce-of-the-future/

Walter, Y. (2024). Embracing the future of Artificial Intelligence in the Classroom: the relevance of AI literacy, prompt engineering, and Critical Thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21, 15. https://doi.org/10.1186/s41239-024-00448-3

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Appendix

Appendix I: The Study Sample
Appendix 2: Coding Scheme
Appendix 3: Coding Concept Distribution and Applications Across Four AI Literacy Levels

Information Research - Vol. 30 No. iConf (2025)

Readers' advisory guide development for webtoon

Hyerim Cho, Denice Adkins, Hillary Gould, Katy Bond, and Ky Flood

DOI: https://doi.org/10.47989/ir30iConf47179

Abstract

Introduction. Readers' advisory (RA) services are one of the core information services that libraries provide. As the types and formats of the collection materials have diversified, relevant RA services have evolved accordingly. This project develops an RA guide for webtoons, emerging online webcomics that have successfully engaged international audiences with their diverse themes and accessible content.

Methods. The current study compiled the data from focus groups, semi-structured interviews, and a large-scale online survey with avid webtoon readers. Then, a webtoon expert research team reviewed the major themes identified in the empirical data and created recommendation categories accordingly. Lastly, subject matter expert consultations were performed to revise the recommendation categories.

Results. The finalized webtoon RA guide has been developed and uploaded to the institutional repository, so it is publicly available to anyone, especially librarians who are interested in providing relevant reference and recommendation services. The developed RA recommendation categories contain nine primary subjects, including art style, mood, cross-media, and representation.

Conclusion. By including accessible digital content in reference services, libraries will be able to expand their reference services beyond the physical collections and promote diverse content to their patrons.

Introduction

Webtoon is a form of online webcomic where creators share their visual storytelling sequentially in a vertical format. Originating from South Korea, webtoons.com, a major webtoon provider, was launched in the United States in 2014. Since its launch, it has rapidly grown in popularity among global audiences. As of September 2023, there were 175 million monthly active users of the platform, and over 12 million professional and amateur creators have shared their works (Webtoon, n.d). Younger adult readers, such as Gen Z (born between 1997-2012) and millennials (born between 1981-1996), have been the primary force driving readership in this new platform, including young female readers who have been neglected in the previous comics industry (Gustines & Stevens, 2022).

Existing webtoons are mostly free to read as long as one has internet connectivity. Their web-based and mobile-based application services are freely provided to anyone, though users can buy advanced access to newly released episodes. Webtoon readers can enjoy their content anywhere, anytime, with the access of their smartphones. With its convenient and accessible format, webtoon supports diverse readers' needs, including readers with reading difficulties, limited time and resources, and the need to find relatable communities (Cho et al., 2022).

As younger readers shift to digital reading platforms like webtoon, it is important for information professionals to understand these newly emerging materials and prepare for future information services that are affected by new reading needs. However, as webtoon is relatively a new, born-digital format, the research to support information professionals in this regard is still lacking.

The current research aims to solve this problem by creating a readers' advisory (RA) guide for webtoons that is supported evidentially by the investigators' recent empirical studies on webtoon readers, coupled with consultations from the subject expert reference librarians. The project's immediate impact on reader advisory practice will be to 1) promote the newly emerging, accessible graphic medium; 2) familiarize librarians with the diverse materials and topics included in webtoon; and 3) support public librarians' reference services by providing an RA guide for webtoon. Potential future impacts will be providing a framework for librarians and information professionals to explore other emerging media formats. The current research answers the following research question:

- **RQ:** What reading needs do webtoon readers have when looking for a new webtoon, and how can the Readers' Advisory guide support these needs?

Literature review

Webtoon is a new storytelling medium, and scholarly research into webtoon is likewise still in its infancy. Existing studies have examined the cultural and artistic aspects of webtoon, focusing on its transnational and transmedia characteristics (Jin & Yi, 2020; Park et al., 2019) and cultural influences on international audiences (Jang & Song, 2017), as many webtoon works originated from South Korea. Several efforts have been made to understand webtoon's roles in literacy education, as well, such as Kim's (2023) study on K-Pop fans' language learning and literacies with webtoon.

RA services connect readers with materials that meet their needs (Watson, 2000), and they have been one of the core information services libraries provide. While '*traditional*' book-based RA services and relevant studies have evolved within LIS, research on RA services for emerging media and multimedia resources, such as webtoon, has not received enough attention within the LIS scholarly communities. Considering library users' born-digital material needs, as well as libraries' shift to digital services (Ex Libris, 2022), the lack of understanding of digital media and preparations for the relevant services is a problem.

Generally, RA services are developed based on the philosophy that reading has an intrinsic value and provides a specific pleasure unachievable in other ways (McCook, 1993). Van Fleet (2003) states that popular and genre fiction is valuable because it appeals to readers and should, therefore, be recommended in library services. These collections serve the needs of scholars, students, and recreational readers through the interaction of reader and text (Fleet, 2003). On a similar note, Smith and Tarulli (2015) argue that the basis of a reader's advisory services is to listen to the person and discern their likes and dislikes in literature.

Existing literature emphasizes the need to understand readers and that RA services should be based on what readers want, but current RA guides and their development often depend heavily on individual experts' subject knowledge. Nonetheless, fiction and popular culture fans have been actively developing their own recommendation and organization systems, such as Archive of Our Own (AO3), a '*non-profit open-source repository for fan-fiction and other fan-works*' (Lorigados, 2022, p.2). Noting the readers' knowledge and reading experiences, the current study develops and presents an RA guide for webtoons from the reader-centred perspective, based on the comprehensive reviews of webtoon readers' reading motivations, appeals, needs, and reading behaviours.

Methods

The current project takes a qualitative and empirical approach to the design of RA. Based on multiple sources of information from avid webtoon readers (focus groups, semi-structured interviews, and a large-scale online survey), data were inductively arranged into broad categories and deductively analysed with regard to individual webtoon titles. This led to the iterative creation, review, and revision of the webtoon guide through group discussions and with the advice of subject matter experts. The current manuscript focuses on the creation of the webtoon RA guide.

Previous phases of this study identified webtoon readers' appeal factors and their reading behaviours. These included format-specific elements such as illustration style, as well as content-specific elements such as subject, mood, characters, or use of specific tropes (Cho et al., 2022). The goal of the current project was to make that information usable for librarians by identifying specific appeal factors and other relevant categories in specific webtoon titles.

Between August and September 2023, a research team was composed of members with webtoon expertise. The data collected from the previous phases were compiled and reviewed by the entire research team to identify the most significant categories for webtoon readers to support their reading needs. Fiction classification is notoriously problematic (Beghtol, 1990). The webtoon platform already identifies several traditional access points, such as genre, author, and series. Following Almeida and Gnoli (2021), the research team chose to focus on webtoon attributes rather than subjects or genres. In addition, the research team used the words and phrases of webtoon readers to identify those attributes as a starting point for the analysis.

After the creation of several overarching RA categories, such as Complexity, Mood, Representation, and art style, four members of the research team engaged in extensive webtoon reading to understand the latest webtoon trends additionally and describe how their webtoon titles fit into those categories by creating sub-categories. For instance, the mood of one webtoon title might be light-hearted, while another might be dark and foreboding. Categories and sub-categories were reviewed and discussed weekly in a whole group setting, wherein researchers explained how and why they designated certain webtoons in different categories. In the context of this discussion, several categories were iteratively revised and refined. This process of creating the initial webtoon RA categories and identifying the appropriate webtoon titles took approximately six months.

As the next step, two subject matter experts from the American library association's (ALA) Graphic Novels and Comics Round Table (GNCRT) provided expert evaluation of the initial RA guide.

Additional feedback from this process was discussed by the researchers and incorporated into the RA guide design, resulting in the final set of revised categories and RA guide.

While the current study adopted several processes in taxonomy design, the list of terms is not closed, and these terms are not mutually exclusive. The research team expects this guide to evolve continuously as webtoon readers and media evolve.

Findings and discussions

Recommendation categories

Table 1 presents the developed webtoon RA guide recommendation categories. Figure 1 is a screenshot of a page from the finalized RA guide. There are nine first-level categories, each including second-level categories. Three second-level categories (romantic interests, disabilities, and fictional + historical) include third-level categories for further detailed recommendations. This guide does not include traditional access points (e.g., genres). Instead, using data from webtoon readers, the research team focused on why readers currently read what they read, resulting in categories highlighting these readers' motivations and needs.

First level category	Second level category	Third level category
Art Style	Early comic book	*
	Experimental	*
	Manga/Anime	*
	Simplified	*
Characters	Heroic qualities	*
	Lovable and cute	*
	Magical and superhuman powers	*
	Powerful women	*
	Quirky and funny	*
	Relatable to adulthood	*
	Villain-centric	*
Complexity	Episodic	*
	Long-running	*
Cross-media	Books/Web novels	*
	Graphic novel available	*
	Movies	*
	TV shows	*
How the story unfolds	Avoiding bad ending flags	*
	Diving into social issues	*
	Non-romantic relationship-	*
	Overcoming challenges and	*
	Power fantasy	*
	Romantic interests	Arranged or political marriages
		Friends to lovers
		Love Triangles
		Opposites attract/ enemies to lovers
		Rags-to-riches romances
	Second chance at life	*
	Underdogs never lose	*
Mood	Lighthearted, relaxing, cozy	*
	Escapist, exciting, immersive	*
	Tense, dark, suspenseful	*
	Heartwarming, healing, hopeful	*
	Scary, dangerous, dystopian	*
Multimedia	Background music	*
	Sound effects	*
	Animation/special effects	*
Representation	BIPOC	*
	LGBTQIA+	*
	Disabilities	Visible disabilities
		Invisible disabilities
Where the story unfolds	Another/parallel/fantasy world	*
	Mythologies/legends reimagined	*
	Fictional + Historical	Eastern aesthetics
		Western aesthetics
	What happens in schools	*
	Learning about different cultures	*
	Work and life	*

Table 1. Webtoon readers' advisory (RA) recommendation categories

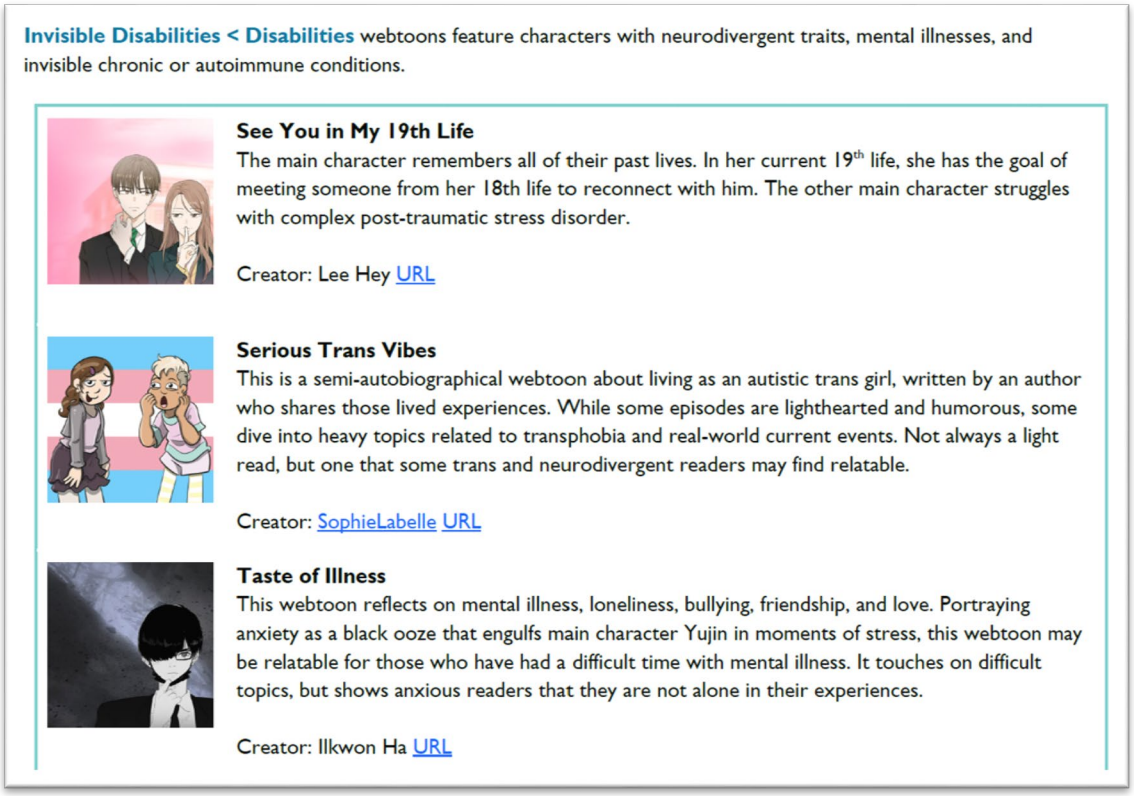


Figure 1. An example page of the finalized webtoon RA guide

For art style, Characters, and Mood, the research team developed second-level categories based on the reviews of the available webtoon titles as well as needs identified from the semi-structured interviews. For the Mood category, particularly, the research team additionally consulted Cho et al.'s (2023) categorizations of moods in the fiction genre to finalize the second-level categories that are consistent with the existing literature.

Many readers mentioned that they like reading webtoons for a short break and prefer works that do not require a lot of mental investment to keep track of the overall story. In contrast, some readers preferred webtoon titles with many episodes published and new episodes still coming out regularly. Due to this reading need, the Complexity category was identified, which consists of episodic and long running. The Episodic category generally does not require any prior knowledge of the webtoon work to begin reading, can be read out-of-order, or picked up at any point in the work's run. Long-running titles (with over 100 episodes) have stronger internal continuity and may require readers to start reading at the beginning of the webtoon to understand the plot.

Webtoon's immersive multimedia elements in reading engagement, such as sound and animation effects (Cho et al., 2022) and its cross-media appeals (MacDonald, 2023) have recently been highlighted. The developed RA categories include cross-media to indicate webtoon titles that are associated with different media formats, such as TV shows or movies, and multimedia to refer to works that incorporate multimedia elements for more immersive reading engagements.

The following section shares some additional challenges and opportunities encountered in the current webtoon RA development projects.

Challenges in granularity: plot/narratives, settings, and representations

The goal of the webtoon RA guide is not to replicate traditional categories of genre or provide a comprehensive list of them, which would be unwieldy to navigate. The broad category headings of

How the story unfolds, Where the story unfolds, and Representation were selected to encompass categories that respond to readers' needs as indicated by previous phases of this study and by popular throughlines in webtoon content and readership. For example, Representation was developed for readers being drawn towards webtoons, which involve characters they could see themselves in or communities that they were a part of (Cho et al., 2022), or for readers who are interested in understanding others. Creating the overarching category of Representation with subcategories addressing under-represented groups supports this need.

In How the story unfolds, the research team recognized the importance of the romance genre (Ramsdell, 2012; Saricks, 2001). However, the researchers also noted that while finding a romance story on webtoon platforms is easy, it is difficult to find a story that features family (or chosen family) relationships or friendships rather than a romantic relationship. Thus, the team decided to add the Non-romantic relationship-centred category for diverse recommendations. However, this decision leads to a discussion of challenges that the research team encountered: when developing an RA guide and recommendation categories, what is the appropriate level of granularity for a category, considering they are not meant to be an exhaustive organization schema but to be a starting point for good advisory services?

As another challenging example, the Representation category includes BIPOC, LGBTQIA+, and Disability, and the Disability category includes sub-categories, Visible and Invisible disabilities (e.g., neurodiversity, mental illness, chronic and autoimmune conditions). For Disability categories, members of these groups were consulted on the wording and how their experiences differ, which led to the necessity of separating these two groups in representation. Initial categories for Representation were at a more granular level, including sub-categories that identify different ethnicities, races, and more. Later, based on the iterative internal discussions and consultations with the subject expert librarians, the current categories were finalized at a less granular level due to the intersectionality of webtoons. Many webtoon titles tend to be diverse and include multiple themes and characters that intersect among these categories; thus, placing a title in one category over the other may result in subjective inconsistencies. Still, it may be fruitful for future studies to investigate whether a certain level of granularity in the RA guide can be more helpful in supporting RA librarians.

Conclusion

The current study developed an RA guide for webtoons as part of a larger research project to understand webtoon readers' motivations, appeals, needs, and reading behavior. The developed RA categories reflect the webtoon titles that are currently in demand and available. Categories such as Characters and Where the story unfolds strongly represent what webtoon readers are looking for now. This indicates that the suggested RA categories can and should evolve. What readers look for might change as time passes. RA services should be prepared for those changes, and so should RA tools and resources. The categories used in this paper represent the words and phrases used by webtoon readers themselves. Future research could address additional ways to incorporate user-generated categories into traditional RA services, moving beyond folksonomies to a wider integration into advisory services for multiple genres and formats.

This research does not apply only to webtoons. Reading materials, and more broadly, types of information objects in general, have evolved in new ways as their formats and platforms have diversified. This indicates that information services expected from libraries will significantly change, too. The shift may have already started; more reference services are conducted online, and more libraries provide born-digital collection access to their users, including streaming services or video games. Considering the inclusion of born-digital materials in library services that can be accessed for free not only helps the library's collection development budget and space concerns but also broadens library users' reading horizons. As media types expand, similar

discussions of appeal factors or recommendations can help library RA personnel learn how users interact with and value the formats and types of media they use.

The full version of the webtoon RA guide, including the definitions of each category and the relevant recommendations for each category, can be found at the University of Missouri's open digital repository for public use (https://mospace.umsystem.edu/xmlui/handle/10355/102741).

Acknowledgments

The current project was supported by the American Library Association's Carnegie-Whitney Grant.

About the authors

Hyerim Cho is an Assistant Professor at the School of Information Science and Learning Technologies, University of Missouri. Her primary research interests are the information needs and behavior of multimedia and popular cultural materials users. They can be contacted at hyerimcho@missouri.edu.

Denice Adkins is Professor at the School of Information Science and Learning Technologies, University of Missouri. Her primary research interests are public libraries and services to diverse communities. They can be contacted at adkinsd@missouri.edu.

Hillary Gould is a PhD candidate in Information Science and Learning Technologies at the University of Missouri whose primary focus is researching and designing educational video games. They can be contacted at hgould@missouri.edu.

Katy Bond is a master's student in Library and Information Science at the School of Information Science and Learning Technologies, University of Missouri. They can be contacted at kebbkd@missouri.edu.

Ky Flood is a master's student in Library and Information Science at the School of Information Science and Learning Technologies, University of Missouri. They can be contacted at kfpb7@missouri.edu.

References

Almeida, P. de, & Gnoli, C. (2021). Fiction in a phenomenon-based classification. *Cataloging & Classification Quarterly*, 59(5), 477–491. https://doi.org/10.1080/01639374.2021.1946232

Beghtol, C. (1990). Access to fiction: A problem in classification theory and practice. Pt. II. *Knowledge Organization*, 17(1), 21-27.

Cho, H., Adkins, D., & Pham, M. (2022). 'I only wish that I had had that growing up': Understanding Webtoon's appeals and characteristics as an emerging reading platform. In *Proceedings of the Association for Information Science and Technology* (ASIS&T), 59(1), 44-54. https://doi.org/10.1002/pra2.603

Cho, H., Lee, W.-C., Huang, L.*, & Kohlburn, J. (2023). User-centered categorization of mood in fiction. *Journal of Documentation*, 79(3), 567-588. https://doi.org/10.1108/JD-03-2022-0071

Ex Libris, A ProQuest Co. (2022, January 3). Three trends shaping the future of libraries. *Library Journal*. https://www.libraryjournal.com/story/three-trends-future

Gustines, G. G., & Stevens, M. (2022, July 23). Comics that read top to bottom are bringing in new readers. *The New York Times*. https://www.nytimes.com/2022/07/23/arts/digital-comics-new-readers.html

Jang, W., & Song, J. E. (2017). Webtoon as a new Korean wave in the process of glocalizaion. *Kritika Kultura*, (29).

Jin, D. Y., & Yi, H. (2020). Transnationality of popular culture in the Korean Wave. *Korea Journal*, 60(1), 5-16.

Kim, H. (2023). 'I read webtoon every day!': young adult k-pop fans' language learning and literacies with Korean webcomics. *Journal of Graphic Novels and Comics*, 14(1), 104-118.

Lorigados, E. B. (2022). *Fan-Fiction and AO3 Free-Form Tagging Practice: Innovating Open-Source Tools for Tag Network Visualization and Analysis*. [Master's thesis, York University].

MacDonald, J. (2023, September 14). Korean dramas play a role in the booming popularity of Webtoons. *Forbes*. https://www.forbes.com/sites/joanmacdonald/2023/09/14/korean-dramas-play-a-role-in-the-booming-popularity-of-webtoons/?sh=7fc4b5a84596

McCook, K. (1993). Considerations of theoretical bases for readers' advisory services. *Collection Building*, 12, 7-12. https://doi.org/10.1108/EB023336

Park, J. H., Lee, J., & Lee, Y. (2019). Do webtoon-based TV dramas represent transmedia storytelling? Industrial factors leading to webtoon-based TV dramas. *International Journal of Communication* (19328036), 13.

Ramsdell, K. (2012). *Romance fiction: A guide to the genre*. Bloomsbury Publishing USA.

Saricks, J. G. (2001). *The readers' advisory guide to genre fiction*. American Library Association.

Smith, D., & Tarulli, L. (2015). Readers' advisory: The who, the how, and the why. *Reference and User Services Quarterly*, 54(4), 11-16. https://www.jstor.org/stable/10.2307/refuseserq.54.4.11

Van Fleet, C. (2003). Popular fiction collections in academic and public libraries. *The Acquisitions Librarian*, 15, 63-85. https://doi.org/10.1300/J101V15N29_07

Watson, D. (2000). Time to turn the page: Library education for readers' advisory services. *Reference and User Services Quarterly*, 40, 143-146.

Webtoon. (n.d.). About. Webtoon. https://about.webtoon.com/

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Exploring the psychological mechanisms of library anxiety considering achievement motivation and goal orientation

Masaki Takeda and Satoru Suto

DOI: https://doi.org/10.47989/ir30iConf47176

Abstract

Introduction. Although library anxiety is recognised as having a negative impact on people's exploration, its mechanisms are unknown. Moreover, no studies have systematically demonstrated the mechanisms through which exploration is modified in the context of library anxiety. This study reviewed the context and results related to library anxiety and hypothesised that achievement motives and goal orientations underlie changes in exploration.

Method. A survey of approximately 2,000 participants was conducted and 642 responses were collected.

Results. A covariance structure analysis of the responses revealed that, in most cases, the hope of success was significantly related to mastery goals and information seeking from all sources, with the aim of personal growth and a deeper understanding of the task. When motivated by fear of failure, participants avoid damaging their self-esteem by envisioning the event of failure or being treated lightly by others.

Conclusions. These results imply that when one's goal is to prevent the loss of self-esteem or being treated lightly by others, seeking information from other people is avoided in favour of searching for information from non-human sources.

Background

Library users are overwhelmed by the complexities of library services and the sheer volume of information sources. Hence, Mellon (1986) investigated the process of writing reports for assignments at the library. Students expressed fear of approaching librarians and their discomfort with the library rather than difficulties in finding information. Users with library anxiety feel helpless when searching for information in the library; they cannot navigate the library well, their library skills are inferior to those of their peers, they do not know how to start searching for materials, and they struggle to find the materials they need. Mellon (1986) states that these reactions are natural and that library anxiety can be resolved by building on successful experiences. Library anxiety has also been examined among postgraduates (Jiao & Onwuegbuzie, 1998) and immigrants (Mengxiong, 1995). Indeed, library anxiety contributes to poor academic performance and needs to be addressed (Jan, Anwar, & Warraich, 2020; Kwon, 2008).

Library anxiety has over time grown to include not only searching libraries, but also the Web. The adoption of digital environments (e.g., digitisation, the web) in libraries has led to widespread library anxiety. For instance, online public access catalogues and other web-based information retrieval services are no longer mediated by librarians and other professionals but are directly provided to users (Bandura, Freeman, & Lightsey, 1999; Mizrachi & Snunith, 2004; Omran, 2001). Users are expected to obtain the information they seek using these technologies.

Library anxiety is primarily induced by users' psychology (Agalya, Singson, Thiyagarajan, & Gogoi, 2022; Akbarzadeh & Bigdeli, 2020; Jan, Anwar, & Warraich, 2020; Kwon, 2008). Jan et al. (2020) focused on emotional intelligence (EI), that is, one's ability to monitor their own and others' emotions, identify them, and adjust one's thoughts and behaviours accordingly. They found a positive relationship between EI and library anxiety (r (140) =0.18, p=0.04), with a one-unit increase in EI inducing a 0.23-unit increase in library anxiety. Thus, students with high EI are sometimes unwilling to utilise libraries due to fear, that is, they experience library anxiety. It is important to note that EI affects how university students (participants) deal with library anxiety and does not directly relate to library anxiety. Jan et al. (2020) showed that psychological factors regulate library anxiety. However, the specific psychological factors EI monitors and their psychological mechanisms have not been clarified regarding library anxiety. Other studies exploring library anxiety-related psychological factors (Agalya et al., 2022; Akbarzadeh & Bigdeli, 2020; Kwon, 2008) have focused on the degree of anxiety experienced and familiarity with libraries and librarians, neglecting not only the multiple psychological factors related to library anxiety but also the psychological mechanisms through which library anxiety occurs. Additionally, differences between searching and stopping searches among those with library anxiety have not yet been addressed.

Most previous studies on library anxiety have addressed exploration as a component of task performance or learning activities. Therefore, based on findings in educational psychology, achievement motivation and goal orientation are psychological factors that may be related to library anxiety. This makes sense as library anxiety is a psychological state that can occur during learning when achievement motivation and goal orientation influence the desire to complete learning tasks and not fail. Achievement motivation is a psychological inclination that can be positively or negatively driven. Positive achievement motivation might be the hope of success, which represents the desire to achieve success in a goal or task. Negative achievement motivation could take the form of fear of failure, which represents the desire to avoid failure. Hope of success is related to experiencing pride when achieving a goal, whereas fear of failure is related to experiencing shame when a goal is not achieved (Elliot & Church, 1997). Goal orientation is a goal that is set and positioned to recognise or monitor one's achievement motivation and satisfy it (Atkinson, 1964). Goal orientations can be categorised into three types: mastery, performance-approach, and performance-avoidance goal orientations (Elliot & Harackiewicz, 1996). Mastery goal orientation refers to being motivated to improve or master a task. Performance-approach

orientation refers to being motivated by the desire to obtain a positive evaluation in a task, whereas a performance-avoidance goal orientation refers to being motivated by fear of failure stemming from being evaluated negatively and being seen as incompetent. Regarding the relationship between these three goal orientations, academic performance, and motivation, several studies have confirmed a positive correlation between performance-approach goals with intrinsic motives and grades and a negative correlation between performance-avoidance goals with intrinsic motivation and grades (Elliott & Dweck, 1998; Elliot & Harackiewicz, 1996; Tanaka & Yamauchi, 2000).

Hypothesis

Many studies on library anxiety have used exploration as part of a task or task performance with learning as the research context. It has been proposed that library anxiety is an effect of achievement motivation and goal orientation—concepts widely studied in educational psychology—and is reflected in factors such as underestimating one's own skills, avoiding behaviours for fear others would see one as incompetent, and changing one's behaviour.

Therefore, in the present study, based on a review of the seminal papers referenced above, it was hypothesised that (1) achievement motivation, which is the inherent psychological inclination of a person, influences goal orientation based on how they approach and perform each task (Atkinson, 1964; Elliot & Church, 1997; Elliott & Dweck, 1998; Elliot & Harackiewicz, 1996; Tanaka & Yamauchi, 2000), and (2) the interaction between achievement motivation and goal orientation modifies exploratory behaviour (Bandura, Freeman, & Lightsey, 1999; Jan, Anwar, & Warraich, 2020; Jiao & Onwuegbuzie, 1998; Kwon, 2008; Mellon, 1986; Mengxiong, 1995; Mizrachi & Snunith, 2004; Omran, 2001).

Methods

Survey participants and procedures

The study participants were all first-year undergraduate students (2,032 individuals) at a Japanese national university. The university has faculties of informatics, engineering, science, agriculture, humanities and social sciences, and education. The survey was conducted between 6-22 July 2021. We limited the survey to first-year undergraduates because we thought that, as they had only recently enrolled, they would all have similar familiarity with the library and classes.

Survey scales

The survey included the revised 10-item version of the Achievement Motives Scale (AMS-R) by Lang and Fries (2006) to measure achievement motives, Elliot and Church's (1997) Achievement Goal Scale as a measure of goal orientation, and the information-seeking behaviour scale by Igarashi (2003).

The AMS-R consists of two subscales: hope of success and fear of failure. Responses to the AMS-R were obtained using a 4-point Likert scale (1=not at all true; 2=not very true; 3=somewhat true; 4=very true). The goal orientation scale consists of three subscales: mastery goals, performance-approach goals, and performance-avoidance goals, while the information-seeking behaviour scale consists of three subscales: implementing mass media resources, implementing interpersonal resources, and independence in information seeking. Both scales were answered using the same 4-point Likert scale employed for the AMS-R. Each scale was translated and survey participants completed the scale in their preferred language. The wording of some items was modified to suit the Japanese learning environment. The achievement motive and goal orientation scales explore general psychological factors that govern behaviour during learning. Although not limited to libraries, they are useful for understanding library anxiety in the context of learning. Participants' tendency to explore was measured using the information-seeking behaviour scale.

Results

Factor analysis, correlation analysis

Overall, 642 (31.59%) responses were submitted: sex, 344 females and 298 males; mean age: 18.31 years (*SD*=1.13).

Tables 1–3 present the results of factor analyses (maximum likelihood method, Promax rotation) for the achievement motive, goal orientation, and information-seeking behaviour scales, respectively. For the AMS-R in Table 1, based on the parallel analysis and factor loadings, a two-factor structure similar to that in previous studies was judged to be optimal: RMSR=.05, RMSEA=.07, ω=.93; additionally, the fit and stability of the model were judged to be sufficiently secure to withstand analysis. The factors were named fear of failure (FF) and hope of success (HS) with reference to previous studies. As in past studies, the goal orientation scale (in Table 2) was deemed ideal, with a three-factor goal structure: mastery goal (MG), performance-approach goal (AP), and performance-avoidance goal (AV) (RMSR= 04, RMSEA=.06, ω=.94). Finally, the previous study's three-factor structure was judged to be optimal for the preference scale linked to information-seeking behaviour (in Table 3) (RMSR=.04, RMSEA=.06, ω=.94) and the factors were named implementing mass media resource (MM), implementing interpersonal resource (IP), independence in Information seeking (II), as per previous studies.

	F1	F2	commonality
Fear of failure			
If you are not sure that you will succeed, you are anxious about everything you do	**.77**	-.05	.59
Feel quite anxious in new situations, even if no one is watching	**.75**	-.06	.55
Fear of failing in slightly difficult situations when so many things are up to	**.73**	-.04	.53
If the problem cannot be resolved immediately, I begin to feel insecure	**.65**	.07	.43
Hope of success			
Interested in situations where they can test their competence	-.06	**0.8**	.67
Attracted by challenges that allow them to test their own abilities	-.04	**0.78**	.61
I like situations where I can discover how much I am capable of doing	-.05	**0.61**	.37
Enjoy situations where you can make the most of your abilities	-.05	**0.41**	.17
When faced with a solvable problem, they are inclined to start solving it	.12	**0.3**	.08

Table 1. Factor analysis: achievement motive scale

	F1	F2	F3	commonality
Performance-approach goal				
I aim to get better grades than others	**.79**	.00	-.01	.62
For me it is important to be seen as being able to do better than everyone	**.75**	-.02	-.15	.49
Getting better marks than others is important to me	**.74**	-.01	.04	.58
I want to get good grades so that my family, friends and others think I can	**.71**	-.11	.05	.50
The thought of getting better grades than everyone else around you	**.61**	.19	-.06	.45
Striving to show that they are more capable than others	**.58**	.19	-.04	.42
I'd like to make sure I don't get worse grades than the rest of you	**.50**	.00	.30	.46
Mastery goal				
I sometimes wish I had a broader or deeper knowledge of the content of	-.04	**.70**	-.01	.48
I prefer to learn new things, even if they are a bit more difficult	.08	**.67**	-.13	.49
I want to learn as much as I can in class	.00	**.60**	.01	.36
I prefer to do interesting content in class, even if it is a little difficult,	.04	**.53**	-.12	.30
I would like to fully understand the material distributed in class	.06	**.50**	.07	.28
It's important for me to make sure that I understand the lesson content as	-.06	**.40**	.26	.21
Performance-avoidance goal				
I would worry about the possibility of getting a bad grade in class	-.12	.10	**.89**	.72
Often wonder what they would do if they got a bad grade	-.14	.10	**.88**	.70
I wish I didn't get bad grades	.05	-.08	**.41**	.19
I don't want to ask strange questions to teachers and TAs and have them	.19	-.13	**.30**	.17

Table 2. Factor analysis: goal orientation scale

	F1	F2	F3	commonality
Implementing interpersonal resource				
When doing something new, ask someone who knows about it	**.79**	-.13	.03	.57
Actively ask others when researching matters of interest	**.78**	-.07	-.11	.54
When in doubt, ask someone you trust	**.68**	-.09	.04	.43
I try to have a wide range of contacts with people who have different types	**.50**	.03	.01	.27
When gathering information, do not ask someone	**.46**	.01	-.20	.20
Independence in Information seeking				
Gather the information you think you need on your own	-.15	**.95**	-.20	.62
Gathering your own information on matters of interest	-.02	**.69**	-.08	.40
I rarely willingly look something up	-.18	**.55**	-.05	.23
I want to get information promptly	.13	**.46**	.02	.29
If I don't understand something, I am quick to look it up	-.04	**.44**	.19	.33
I want to investigate thoroughly until I am satisfied	.07	**.41**	.19	.25
I prefer to know a wide variety of things	.10	**.37**	.09	.15
As much as possible, I want to know the essence of things	.19	**.30**	.19	.22
Implementing mass media resource				
If you don't understand something, first look it up in a book, magazine or	-.17	-.11	**.92**	.66
When you want to know something, look at books, magazines and websites	-.11	.01	**.72**	.49
Refer to books, magazines and websites when doing something new	-.06	-.06	**.67**	.38

Table 3. Factor analysis: information-seeking behaviour scale

Table 4 presents the correlation coefficients of each subscale, illustrating that FF was strongly and positively correlated with AV (r=.54, p<.01); but not with MG. HS was subsequently found to be strongly positively correlated with MG (r=.48, p<.01) and AP (r=.40, p<.01) but not with AV. These results are consistent with those reported by Elliot and Church (1997). MG was significantly and positively correlated with AP (r=.34, p<.01). For information-seeking behaviour, AV was moderately positively correlated with MM (r=.43, p<.01), and MG with IP (r=.39, p<.01), II (r=.42, p<.01), and MM (r=.41, p<.01). Other positive correlations were found between IP with II (r=.42, p<.01) and MM (r=.34, p<.01), and between II with MM (r=.78, p<.01).

	1	2	3	4	5	6	7	8
1. FF	-							
2. HS	.08†	-						
3. AV	.54**	.08†	-					
4. MG	.04	.48**	.06	-				
5. AP	.16**	.40**	.05	.34**	-			
6. IP	.01	.18**	.05	.39**	.11*	-		
7. II	.01	.27**	.06	.42**.	.16**	.42**.	-	
8. MM	.14**	.29**	.43**.	.41**	.25**	.34**	.78**	-

Table 4. Correlation coefficients between subscales

The impact of achievement motive and goal orientation on information-seeking behaviour

A covariance structure analysis was conducted to test the model in which the two achievement motives led to information-seeking behaviour, either directly or via goal orientation (Fig. 1). The model's goodness of fit was generally good, with GFI = 1.00, AGFI = .91, CFI = .95, and RMSEA = .07.

Looking at the path from achievement motive to goal orientation, HS had a direct positive influence on AP (.39) and MG (.47). FF had a direct positive influence on AV (.54). IP (.40) positively influenced MG, II (.39), and MM (.41). Additionally, AV had a direct positive effect on MM (.41). IP and II were influenced only by MG, whereas MM was influenced by both MG and AV to the same extent. Focusing on indirect effects, HS affected IP (.19), II (.18), and MM (.18) via MG. Conversely, FF was found to affect MM (.22) via AV (.22).

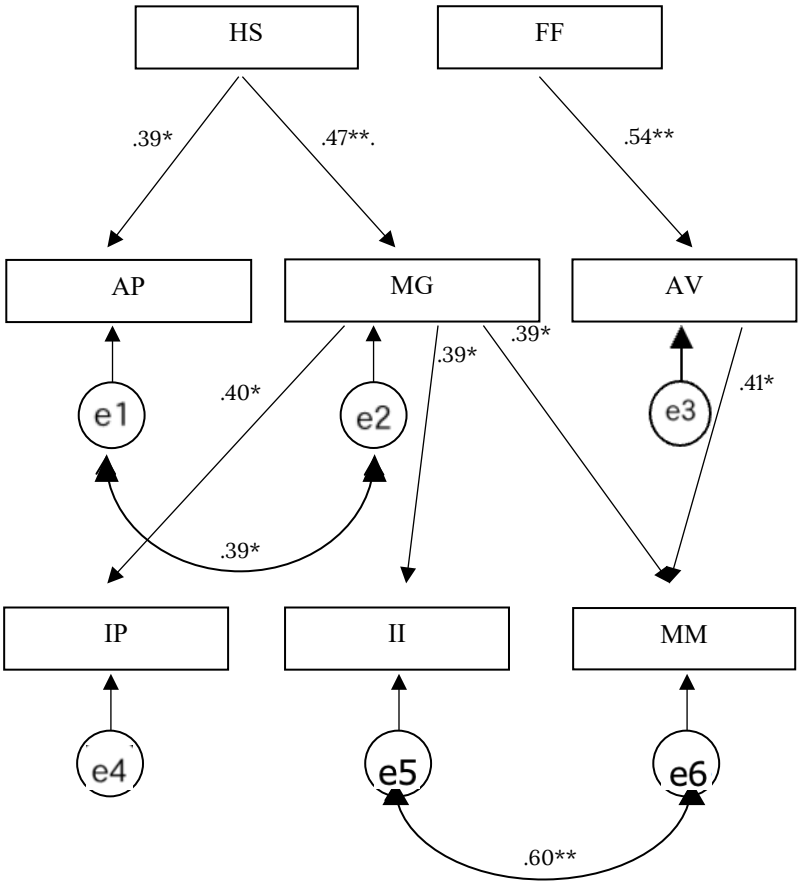


Figure 1. Causal model of achievement motive, goal orientation and preferences in information-seeking behaviour (Only significant paths are shown; * p<.05, ** p<.01)

Discussion

First, regarding hypothesis 1, the results confirmed that hope of success has a direct positive influence on performance-approach and mastery goals and that it significantly influences mastery goals. Further, it was confirmed that fear of failure directly influences performance-avoidance goals; thus, hypothesis 1 was accepted. This result was consistent with the findings of Elliot and Church (1997) and Mitsunami (2010). Hope of success appears to lead to mastery goals, whereas fear of failure tends to lead to performance-avoidance goals. Regarding hypothesis 2, the results showed that the performance-approach goal had no effect on the implementing interpersonal resources, independence in information seeking, or implementing mass media resources, while the mastery goal had a similar positive effect on these three factors, and the performance-avoidance goal had a positive effect on implementing mass media resources; thus, hypothesis 2 was partially supported.

Hope of success affected mastery goals in most cases and had the same effect on implementing interpersonal resources, independence in information seeking, and implementing mass media resources. Therefore, when the focus is on personal growth and a deeper understanding of the task rather than on comparison with others, information seeking is actively carried out using all sources of information. However, when fear of failure is the motive and a performance-avoidance goal is adopted, searching for information through others is avoided, and searching using interpersonal resources is preferred to prevent damage to self-esteem or being treated lightly by others in the event of failure. Mass media resource searches tend to occur regardless of goal orientation.

Limitation and future study

This study was conducted among university students. As attitudes towards tasks and challenges in learning are thought to differ at each educational stage, the results obtained cannot be generalised to all educational stages. In addition, this study adopted a simplified survey and model to first grasp the potential psychological mechanisms involved in library anxiety. However, the fact that the performance-approach goal did not affect any of the exploratory behaviours suggests the existence of other variables. Building on this research, it will be necessary to investigate the relationships between various cognitive factors and learning-related variables.

Acknowledgements

This research was supported by JSPS Grants-in-Aid for Scientific Research (23K17162) and SUNTORY FOUNDATION.

About the authors

Masaki Takeda is Assistant Professor in the Institute of Library, Information and Media Science, University of Tsukuba, Japan. His research focuses on universal services in library. He received his Ph.D. from University of Tsukuba. He can be contacted at masakita@slis.tsukuba.ac.jp

Satoru Suto is Professor in Faculty of Global Interdisciplinary Science and Innovation, Shizuoka University, Japan. He received his Ph.D. from Chuo University and his research interests are in the cognitive psychology. He can be contacted at suto.satoru@shizuoka.ac.jp

References

Agalya, A., Singson, M., Thiyagarajan, S., & Gogoi, T. (2022). Investigating the relationship between emotional intelligence, library anxiety, and academic performance of postgraduate students. Journal of Information and Knowledge, 59(5), 295–296. https://doi.org/10.17821/srels/2022/v59i5/170654

Akbarzadeh, F., & Bigdeli, Z. (2020). Library anxiety of residents in the use of information sources and electronic services based on the Localized Bostick Scale. Payavard Salamat, 13(5), 359–368.

Atkinson, J. W. (1964). An introduction to motive. Van Nostrand.

Bandura, A., Freeman, W. H., & Lightsey, R. (1999). Self-efficacy: The exercise of control. 158–166. https://doi.org/10.1891/0889-8391.13.2.158

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motive. Journal of Personality and Social Psychology, 72(1), 218. https://doi.org/10.1037/0022-3514.72.1.218

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motive: A mediational analysis. Journal of Personality and Social Psychology, 70(3), 461. https://doi.org/10.1037/0022-3514.70.3.461

Elliott, E. S., & Dweck, C. S. (1998). Goals: An approach to motive and achievement. Journal of Personality and Social Psychology, 54(1), 5. https://doi.org/10.1037/0022-3514.54.1.5

Igarashi, T. (2003). Development of the information-exploration scale and its relation to undergraduates' media usage. Educational Sciences, 50, 285–292.

Jan, S. U., Anwar, M. A., & Warraich, N. F. (2020). The relationship between emotional intelligence, library anxiety, and academic achievement among university students. Journal of Librarianship and Information Science, 52(1), 237–248. https://doi.org/10.1177/0961000618790629

Jiao, Q. G., & Onwuegbuzie, A. J. (1998). Perfectionism and library anxiety among graduate students. The Journal of Academic Librarianship, 24(5), 365–371. https://doi.org/10.1016/S0099-1333(98)90073-8

Kwon, N. (2008). A mixed-methods investigation of the relationship between critical thinking and library anxiety among undergraduate students. College & Research Libraries, 69(2), 117–131. https://doi.org/10.5860/crl.69.2.117

Lang, J. W. B., & Fries, S. (2006). A revised 10-item version of the Achievement Motives Scale. European Journal of Psychological Assessment, 22(3), 216–224. https://doi.org/10.1027/1015-5759.22.3.216

Mellon, C. A. (1986). Library anxiety: A grounded theory and its development. College & Research Libraries, 47(2), 160–165. https://doi.org/10.5860/crl_47_02_160

Mengxiong, L. (1995). Ethnicity and information seeking. The Reference Librarian, 23, 49–50, 123–134. https://doi.org/10.1300/J120v23n49_09

Mistunami, M. (2010). Influence of achievement motive and goal orientation on learning behavior: Difference in cognitive strategies. Educational Psychology, 58, 348–360. https://doi.org/10.5926/jjep.58.348

Mizrachi, D., & Snunith, S. (2004). Computer attitudes and library anxiety among undergraduates: A study of Israeli B. Ed students. International Information & Library Review, 36(1), 29–38. https://doi.org/10.1080/10572317.2004.10762621

Omran, A. I. B. (2001). Library anxiety and internet anxiety among graduate students of a major research university. University of Pittsburgh.

Tanaka, A., & Yamauchi, H. (2000). Causal models of achievement motive, goal orientation, intrinsic interest, and academic achievement in the classroom. Japanese Journal of Psychology, 71, 371–324. https://doi.org/10.4992/jjpsy.71.317

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Towards a sustainable automated library operation; the imperatives of digital skills and related factors of digital skills development of university library professionals in Nigeria

Gbenga Adetunla and Gobinda Chowdhury

DOI: https://doi.org/10.47989/ir30iConf47173

Abstract

Introduction. Research shows that libraries in Nigeria are still grappling with automating their library operations because of lack of skilled workforce. This study has explored the current digital skills and training gaps of library professionals in Nigeria.

Method. This study adopted a survey design using a mixed method strategy of both quantitative and qualitative procedures to collect data through questionnaires and interviews sequentially from 453 respondent.

Analysis. Quantitative data was analyzed using SPSS version 23 and qualitative data was subjected to thematic content analysis.

Results. Findings revealed that the university library professionals in Nigeria lack operational digital skills to carry out tasks especially serials, acquisition, cataloging management etc. Also, found that in-house training is the only pathway for expert training on digital operations while others are ineffective. Personal factors such as financial constraints and gender related issues, alongside management factors such as policy and poor financing were found to impede capacity development of library staff in Nigeria.

Conclusion. The study recommends re-harnessing the existing resources and refocusing the training programs towards expert programs.

Introduction

Library automation involves applying technology to manage in-house library operations such as circulation, cataloging, reference, serials, and acquisitions management (Lawal-Solarin et.al. 2018). While many Nigerian university libraries have adopted library management technologies, they still struggle with effective digital library operations, a challenge that became especially evident during the COVID-19 pandemic when all libraries were shut down (Adetunla et.al., 2023). The International Federation of Library Associations (IFLA) recommended that libraries shift to digital services to continue functioning, but Nigerian universities were unable to transition to virtual operations due to resource limitations. These challenges stem from financial constraints, a lack of workforce capacity, and the excessive costs and technical difficulties of acquiring technology, which have hindered the development of digital library services (Ifejeh &Yusuf,2020; Ameh et.al,2021; Corrado,2023).

Despite various efforts since the 1980s, including external interventions like the World Bank and Carnegie/MacArthur automation projects in technical collaboration with the Mortenson Center; University of Illinois till 2015 (Alabi, 1987; Zubair et.al., 2023), many of Nigeria's library automation initiatives have failed as these efforts became unsustainable after the withdrawal of external technical support caused by the expiration of the agreements and universities struggled to maintain these projects due to a lack of staff expertise. The sustainability of digital information services has been discussed in frameworks such as the one by Chowdhury (2014), which focuses on economic, social, and environmental pillars. He elucidates that economic sustainability of digital information services is to ensure cheaper, easier, and better access to information by reducing direct and indirect costs. Social sustainability of digital information services gives equitable access. As shown in Figure 1, for Nigeria, achieving sustainable automated library services has been linked to embracing open-source technologies, promoting open-access information, and focusing on context-specific skills development for library professionals (Karioja, 2013; Anunobi, 2013; Adetunji & Oladokun, 2020; Onwubiko, 2023).

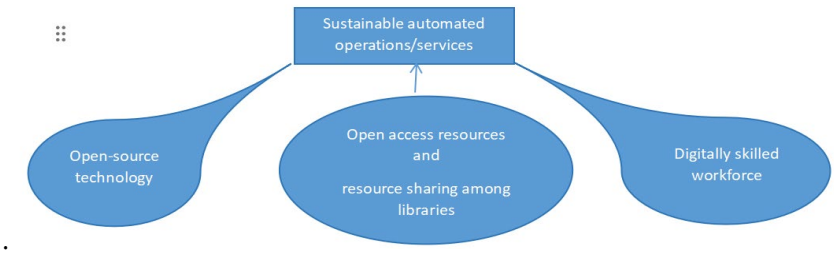


Figure 1. Library automation sustainability chart (self-developed)

The adoption of open-source software, particularly KOHA, has provided hope for automation projects in Nigerian university libraries due to its low cost and ease of maintenance (Tella et.al., 2017; Omopupa et.al, 2020). Additionally, there is growing interest in open access resources, with Nigerian libraries promoting Open Education Resources (OER) in line with the National University Commission's (NUC) mandate for universities to host academic publications on institutional repositories (Corrado,2023; Ogunbodede & Cocodia, 2023; Adedoyin, et.al.,2024). However, despite these advancements, only 5% of Nigerian university libraries have fully automated operations, and the overall state of automation is underwhelming. Research shows that a significant challenge is the lack of a skilled workforce, as 78% of Nigerian information professionals are lacking the digital skills necessary to manage modern technologies (Tella et.al., 2022; Owolabi et.al., 2022). This skills gap was highlighted during the COVID-19 pandemic when Nigerian university libraries, despite investments in technology, struggled to provide services due to the lack of digitally skilled staff (Adetunla et.al., 2023).

The essential digital skills required for library staff to function in this epoch, include basic competencies like online searching, word processing, Wi-fi connection and email services, as well as task-based skills for managing automated tasks such as cataloging, circulation, institutional repositories etc. (Okuongehe & Achugbue 2020; Corrado, 2023). This study aims to investigate the current operational digital skills gap in Nigerian university libraries, explore existing digital skills development programs, and identify staff-side and managerial challenges affecting digital skills development for library professionals. The following research questions were raised:

RQ1. What are the task-based digital skills possessed by library professionals for digital operations in Nigeria?

RQ2. What are the existing training programs on digital operations available to library professionals in Nigeria?

RQ3. What are the staff-side factors and challenges of the existing digital skill development programs on digital operations in Nigeria?

RQ4. What are the managerial factors of the digital skill development and training programs of university library staff in the use of library automation in Nigeria?

Methodology

Research design and sampling technique

This study adopted a survey design using a mixed method strategy of both quantitative and qualitative procedures sequentially. A multi-stage sampling technique was used for this study (see figure 2 below). The first stage involved the use of cluster sampling to put the universities to the 3 geopolitical regions in the South of Nigeria. The second stage involved a random sampling technique to select 3 universities from each region to represent the 3 generations of universities in Nigeria and selection of 3 privately funded universities, and in the third stage, a total enumeration technique was used to sample the 453 professionals (all librarians and library officers) of these 12 universities.

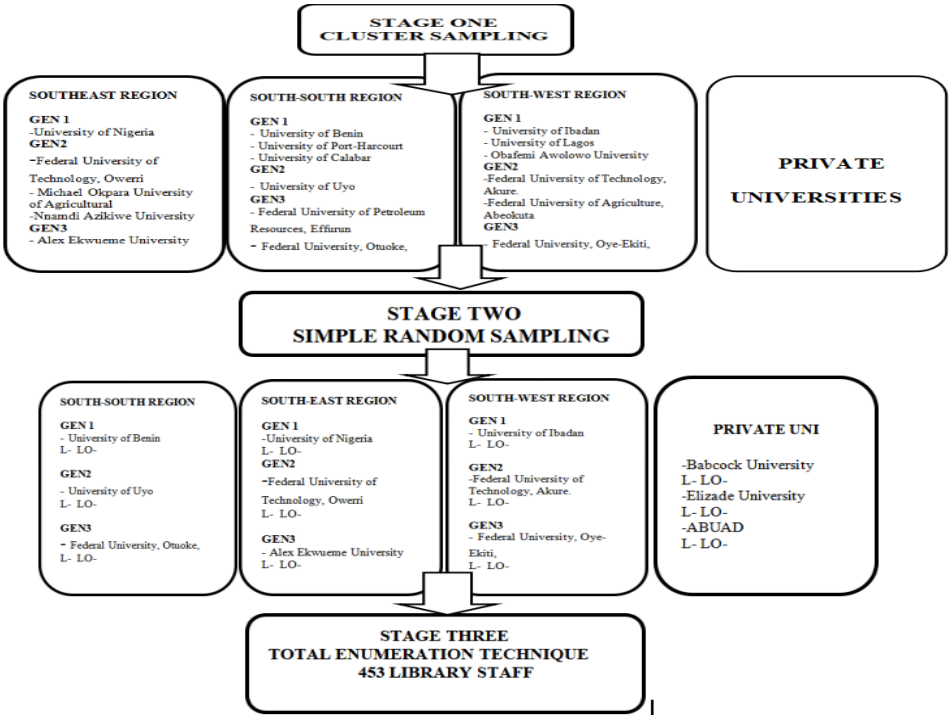


Figure 2. Multistage sampling figure

Data collection and analysis

A mixed method of quantitative and qualitative approach was employed to collect data. Quantitative data was collected using structured questionnaires. 453 questionnaires were directly administered to the respondents across 12 universities in Nigeria. 432 questionnaires were retrieved, and 408 questionnaires were properly filled and found useful for the analysis of this study. Data collected for the study was analyzed using SPSS version 23. In the qualitative phase, data was collected from the senior management staff of the 12 selected university libraries in Nigeria through a digitally recorded interview via zoom. This research set out to interview 12 senior management staff but due to the schedule of the participants, only 10 were interviewed. These 10 senior management staff are within the rank of Principal Librarian to University Librarian. The interview recordings were coded, transcribed, and subjected to thematic content analysis. Research questions 1, 2 and 3 were addressed using the quantitative approach while research question 4 was addressed using qualitative approach.

Presentation of result and discussion

Demographics distribution of respondents

Institution	FUTA	FUTO	UNN	FU'OYE	FU'OTU	AEU	UNI'BEN	UI	UNI'UYO	BABCOCK	ELIZADE	ABUAD	TOTAL
Frequency	17	80	78	10	18	22	32	70	30	25	10	16	408
Percentage	4.2	19.6	19.1	2.5	4.4	5.4	7.8	17.2	7.4	6.1	2.5	3.9	100%
Gender	**male**		**female**		**prefer not to say**								
Frequency	168		226		14								408
Percentage	41.2		55.4		3.4								100%
Designation	**Librarian**			**Library Officer**									
Frequency	164			244									408
Percentage	40.2			59.8									100%
Education	**Diploma**			**Bachelors**			**Masters**			**PhD.**			
Frequency	159			156			65			28			408
Percentage	38.9			39.0			15.9			6.9			100%
Work Experience	**0 – 5**			**6 – 10**			**11 above**						
Frequency	83			202			123						408
Percentage	20.3			39.5			40.1						100%

Table 1. Analysis of demographics distribution of respondents

Table 1 shows the demographic analysis of the respondents. Based on institutions, Federal University of Technology, Owerri has the most respondents while Elizade University has the least respondents. Result based on gender shows that there is more female representation. Analysis of designation shows that library officers are more than librarians. Based on educational qualifications, it was noted that staff that hold Diploma are more, and few PhD holders. It was noted that most of the respondents had between 6 to 10 years' working experience.

Result of quantitative analysis of research questions

Research question 1: what are the task-based digital skills possessed by university library professionals in Nigeria?

OPERATIONAL TASK DIGITAL SKILLS						
Acquisition operation	Manage vendors	Manage budget	Create purchase order	Manage order/ receipt	Handle purchase suggestions	
$\bar{X}$	1.79	1.84	1.82	1.79	1.71	
SD	.578	.607	.719	.661	.746	
Remark	unskilled	unskilled	unskilled	unskilled	unskilled	
Cataloging	Edit bibliographic records	Manage OPAC	Generate call mark/ accessioning	General cataloging	OCLC importing	Z39.50 tool
$\bar{X}$	3.50	3.48	3.45	3.43	1.83	1.75
SD	1.337	1.345	1401	1.418	.684	.728
Remark	skilled	skilled	skilled	skilled	unskilled	unskilled
Serial	Manage subscription	Create new issues/Manage collection	Check subscription expiration	Manage numbering pattern	Claim late issues & manage frequency	
$\bar{X}$	3.43	1.87	3.50	1.91	1.82	
SD	1.416	.650	.1.343	.661	.656	
Remark	skilled	unskilled	skilled	unskilled	unskilled	
Circulation	User registration	Charge & discharge	Renew & prompt due dates	Inter-library operations	Generate reports	Hardware usage
$\bar{X}$	3.37	3.46	3.43	3.47	3.46	3.46
---SD	1.439	1.362	1.400	1.355	1.402	1.366
Remark	skilled	skilled	skilled	skilled	skilled	skilled
Patron	Create new user category	Register new patron profile	Manage/edit patron account	Staff/patron permission		
$\bar{X}$	3.42	3.40	3.43	3.42		
SD	1.408	1.423	1.393	1.391		
Remark	skilled	skilled	skilled	skilled		

Table 2. Analysis of task-based digital skills of university library professionals in Nigeria

Table 2 shows university library professionals in Nigeria lack digital skills for acquisition and serials operation. While they possess some skills for cataloging operation, it is evident they are not skilled in using the Z39.50 tool and OCLC. However, they are skilled in the circulation and patron operation. The extent of automation in the university libraries in Nigeria shows the level of contact the professionals have with the various operations of their library management systems; hence the reason for an average level of skill in cataloging and circulation routine and a low level of skill in the other library routine they are less exposed to such as acquisition and serial routines (Ajani & Buraimo, 2021).

Research question 2: what are the existing training programs on digital library services available to library professionals in Nigeria?

TRAINING PROGRAMS ON DIGITAL LIBRARY SERVICES							
Training programs	**In-house training**	**Workshop & conferences**	**Train-the-trainer**	**Short courses**	**Training tour**	**Online tools**	**Self sponsorship**
$\bar{X}$	3.34	2.34	1.72	1.85	1.64	1.73	1.66
SD	1.498	1.007	.565	.617	.649	.672	.680
Remark	agree	disagree	disagree	disagree	disagree	disagree	disagree

Table 3. Analysis of the training programs on digital operation

Table 3 indicates that most library professionals in Nigerian universities have participated in internally organized training programs on digital library services. These programs are often conducted by vendors when new technologies are deployed in libraries, fulfilling contract obligations (Tella and Akande, 2018; Igbokwe, 2023). However, other training programs were found to be ineffective in enhancing digital skills. Previous research shows that many library professionals in Nigeria have not attended specialized training in digital skills, instead relying on generic conferences, workshops, and courses that do not address the specific expert skills needed for digital library services (Akintola et.al., 2022; Ikolo and Nongo, 2022; Obuh, 2019).

Research question 3: what are the staff-side factors that impede the existing digital skills development of library professionals towards library automation in Nigeria?

Challenges of Digital Skills	**N**	$\bar{X}$	**SD**	**Remark**
Personal factors				
I have phobia for computers and technology	408	1.40	.583	Disagree
My age affects the use of computers and technology	408	1.66	.532	Disagree
My religious belief is averse to the use of computers	408	1.70	.561	Disagree
I have health challenges that requires me to avoid technology	408	1.70	.568	Disagree
I am not financially buoyant to afford training	408	3.86	.621	Agree
My formal LIS training excluded training on digital technology	408	3.65	.624	Agree
Gender related factors				
Family commitments give me little time for training	408	3.67	.649	Agree
My marriage reduces my digital skills training opportunity	408	3.65	.670	Agree
My gender provides little opportunity for digital skills training	408	3.41	1.422	Agree
Organizational factors				
My rank in the office reduces my training opportunities	408	1.66	.672	Disagree
I am rarely promoted and there is no need to attend any digital skills training	408	1.39	.581	Disagree
My library management does not encourage digital skills training	408	3.67	.649	Agree

Table 4. Staff-side factors that impede digital skill development programs

Table 3 highlights several personal and organizational factors that hinder the development of digital skills among library staff in Nigerian universities.

Personal factors: library staff reported that their prior formal library school education lacked courses on digital library technology, creating a significant gap in skills development. Many library schools offer outdated curricula, and even when digital courses are available, they are often

theoretical due to insufficient practical resources and qualified instructors (Ikolo and Nongo, 2022). Additionally, many staff members cannot afford self-sponsored digital training, particularly international programs that require travel and foreign currency, a challenge exacerbated by economic factors in developing countries like Nigeria (Aregbesola et.al, 2019). Gender-related issues also affect women more than men, as care-giving responsibilities (e.g., childbirth and family care) lead to career interruptions and fewer opportunities for digital skills development (Adjah & Walt, 2019; Akhtar & Soroya, 2021).

Organizational factors: on the organizational side, library staff perceive a lack of commitment from university library management to support digital skills development (Akintola et.al., 2022). Financial constraints and low budget allocations for staff development result in limited funding for training programs. The non-payment of salaries also negatively impacts staff morale, discouraging participation in training opportunities. Economic challenges in Nigeria and other developing countries further exacerbate the situation, leaving university management struggling to meet their obligations for staff development and salary payments (Mapulanga,2014; VanguardNewspaper(https://www.vanguardngr.com)).

Result of qualitative analysis of research questions

Research **Question 4:** what are the managerial challenges of the digital skill development and training programs of university library staff in the use of library automation in Nigeria?

Theme 1: policy on digital skills development of university library staff in Nigeria

Arising from the discussion from the senior management staff of the libraries, it was gathered that at the library management level, there is no specified capacity development policy, they operate within the university-wide policy framework which does not take into consideration the context base need of the library staff. The following interview excerpts illustrate this point.

> *The university has a policy on skill training for library staff, but I am not sure there is any policy specifically stating digital skills, but I know, the university has a document on the general training of staff which also includes the library staff (Participant no. 7).*
>
> *There is no strict document on digital skills training, but management deems it fit and necessary to organize training from time to time when the need arises, especially when they look at the trend of things, when there are system upgrades, and it looks as if it is getting difficult for library staff to get along with the operational dimension of that software. Most of the time, they organize training when the need arises (Participant no. 1).*
>
> *Presently there is no writing policy on digital capacity development in my university as at now. But informally, most of the staff in the library are expected to be digitally literate (Participant no. 4).*

Theme 2: funding of training towards digital skills development

In terms of funding sources to the library, the Statutory10% annual allocation from the parent institution, which is supposed to be the major source of funding to the library only exists on the policy document, is not made available to the libraries due to dwindling budgetary allocation to the parent institution. However, the parent institution assumes the responsibility of funding the library for purchases and staff training based on available resources. Furthermore, the libraries benefit from external funding, especially from the federal government. This shortfall in direct funding of library staff for capacity development is sometimes made up for by external bodies such as TETFund (Tertiary Education Trust Fund). TETFund is a federal government establishment

which among other functions, make funds available for capacity development in Nigerian universities. The following interview excerpts illustrate this.

> *In Nigeria, the principle of the statutory 10% annual budgetary allocation from the institution for library operations is only on paper. It is what the usual standard should be, but it is more of a document issue than a reality (Participant no. 3).*
>
> *We usually write to the University management for sponsorship of any library staff going for training (Participant no. 7).*
>
> *My university sends library staff to conferences provided a request for financial commitment was made early enough for the conference and approved. However, the University does not release funds for staff skills development except when it is being pushed by the Librarian, where only selected staff will attend such conferences while others will be in their duty post (Participant no. 9).*
>
> *TETFund has done a lot in the library area of digital skills acquisition. Most of the technological devices in our library were purchased with funds made available by TETFund. TETFund had always contributed to library staff training, digital skills, and upscaling. (Participant no. 5).*

Theme 3: library management efforts toward digital skills development

There is a mixed reaction to effort, as some have taken several initiatives towards training library staff after the COVID-19 pandemic while others remained nonchalant. Libraries that have made concerted efforts to improve the digital skills of library staff did so through internal training approach by bring resource expert to training library staff in form workshops, seminars etc. Most university libraries in Nigeria claim to have the facilities to organize training on digital library operations or services. The availability of internet and computer hardware is also an issue. However, some university managements do not prioritize their library needs when it comes to the issue of training library staff while some library heads do not even make necessary case from the parent institution for training to improve the digital skills of their staff on digital operations or services. The following interview excerpts illustrate this:

> *The Covid era changed library operations. We have done more training and workshops on improving library staff. We are now better than before the pandemic (Participant no. 9).*
>
> *The University management brought some experts for training and seminars to teach and practically demonstrate the use of technology within the library (Participant no. 7).*
>
> *We have some facilities to organize digital training skills but insufficient, like in the aspect of low internet connectivity, to participate in training. Also, hardware is minimal (Participant no. 2).*
>
> *The school management rarely thinks about organizing training for the library staff unless the Librarian pushes it. The only time the management takes the library seriously is when course accreditation is on the radar (Participant no. 10).*

Conclusion and recommendation

University libraries play a vital role in supporting teaching, research, and lifelong learning by providing access to up-to-date scholarly content. In the post-COVID era, libraries have become more digitally focused to meet user needs. However, if Nigerian university libraries fail to adapt towards global trend, the quality of education and research could suffer. A digitally skilled workforce is essential for libraries to meet the demands of the digital age. This study identifies significant gaps in digital skills, particularly in areas of in-house operations like serials, acquisition, cataloging, etc., and highlights the challenges hindering capacity development. These include gender-related factors, lack of expert training programs, and managerial issues such as inadequate policies and funding.

Addressing these challenges will not only improve the quality of the library workforce but also ensure the effective use of technology, positively impacting the quality of education and research in Nigerian universities. Globally, this will contribute to the achievement of the United Nations Sustainable Development Goals 4, 5, 8, 10, and 17. Based on the findings, the study emphasizes the urgent need for a comprehensive digital skills development intervention for library professionals in Nigeria. The following recommendations are made:

Institutional-level action: university libraries should conduct a thorough, context-based analysis of the digital skills needs of their staff and implement targeted training to address gaps. Special support should be provided for female professionals to overcome gender-related barriers to skills development.

Policy development: Nigerian universities without a digital skills development policy for library staff should create one. The policy should define who should be trained, the types and methods of training, and the frequency and location of these training programs.

Focused training programs: university library administrators should prioritize expert-focused training programs that address specific operational needs rather than sending staff to generic, non-specialized training.

Partnerships and funding: libraries should collaborate with internal and external stakeholders, including national and international organizations, to secure funding for training programs that enhance digital skills.

National-level action: professional bodies and stakeholders in the information science sector should develop context-based, timely training plans to address the current digital skills gap in Nigerian libraries.

Acknowledgements

The authors would like to thank the participants of this study for their time and cooperation.

About the authors

Gbenga Adetunla is a Doctoral researcher at the University of Strathclyde, UK. He received his Maters from Babcock University. His research focuses on sustainable digital library operations and can be contacted at gbenga.adetunla@strath.ac.uk

Gobinda Chowdhury is a Professor in Information Science at University of Strathclyde, UK. He received his PhD from University of Sheffield. His research focuses on sustainable digital libraries and information services. He can be contacted at gobinda.chowdhury@strath.ac.uk

References

Adedoyin, O.B., Altinay, F., Gemikonakli, E. *et al.* (2024). National Policy on Open Educational Resources for Higher Education in Nigeria: Evaluation of Institutional Compliance Rate to Infrastructure and the Connectivity Goal. *High Educ Policy* https://doi.org/10.1057/s41307-024-00387-8

Adetunji, O & Oladokun, T (2020). Effective Strategies for Sustainable Library Automation in Nigerian Academic Libraries: Preparing for 'The New Normal' Caused by the COVID-19 Pandemic. *Nigerian Library Association (NLA) Scholarly Conference on Library and Information Services during and after COVID-19 Pandemic*

Adetunla, G., Rasmussen Pennington, D., Chowdhury, G. (2023). Automation of University Library Operations: An Analysis of the COVID-19 Pandemic Experience in the United Kingdom and Nigeria. In: Sserwanga, I., *et al.* Information for a Better World: Normality, Virtuality, Physicality, Inclusivity. iConference 2023. Lecture Notes in Computer Science, vol 13972. Springer, Cham. https://doi.org/10.1007/978-3-031-28032-0_19

Ajani, Y.A., Dunmade, A.O., Tella, A., & Adeniran, C.O. (2023). Information Professionals of the Future and their Prospects in the Era of Fourth Industrial Revolution: The Need for Transformative Potential in Nigeria. Mousaion: South African Journal of Information Studies.

Ajani, F.O. and Buraimo, O. (2021). Perceived impact of automation on university library services by library personnel in Southwest, Nigeria. Information Development. Vol.38(2), 179-191

Akhtar, Gulnaz & Soroya, Muhammad Sahid (2021). Factors Influencing Career Routes of Female Librarianship: A Literature Review. *Library Philosophy and Practice (e-journal).* 6468

Akintola, B.O., Adetomiwa, B., & Onifade, F.N. (2022). Staff Training and Career Development among Library Staff in Federal Universities in Southwest, Nigeria. International Information & Library Review, 54, 315 - 324.

Akpebu Adjah, O., & Van der Walt, T. (2019). Career progression of female librarians in public university libraries in Ghana. Journal of Librarianship and Information Science, 51, 331 - 345.

Ameh, G.J., Ukwuoma, H.C., Oye, P.O. (2021). Covid-19 Pandemic and Evolving Library and information Services: Lessons for Nigeria. *International Journal of Development and Management Review.* Vol. 16(1)

Anunobi, C. (2013), Human Capacity Building in Nigerian University Libraries: An Imperative for Academic Libraries' Contribution towards National Development. *African Journal of Library, Archives, and Information Science* 23(1):33-44

Aregbesola, A., Van der Walt, T, Owolabi, S., Idiegbeyan-ose, J., Okocha F. and Eyiolorunshe, T. (2019). Infopreneurship in a developing country: Opportunities and Challenges. 1st International Conference on Sustainable Infrastructural Development 24–28 June 2019, Covenant University, Canaan Land, Ota, Nigeria. *IOP Conference Series: Materials Science and Engineering*, Vol 640, **DOI** 10.1088/1757-899X/640/1/012123

Chowdhury G. (2014) Sustainability of digital libraries: a conceptual model and a research framework. *International Journal of Digit Library*.14:181–195

Corrado, E. M. (2023). Proprietary and Open Source Software Systems in Libraries: A Few Considerations. *Technical Services Quarterly*, 40(3), 202–209. https://doi.org/10.1080/07317131.2023.2226434

Ifijeh, G & Yusuf, F (2020). Covid–19 pandemics and the future of Nigeria's university system: The quest for libraries' relevance. *The Journal of Academic Librarianship*, 46 (6).

Igbokwe, P C.(2023) Capacity-building practices for librarians' effective service delivery in federal university libraries in Nigeria. Library management, Vol. 44(8/9), p 469 - 483

Ikolo and Nongo, (2022). Training needs of medical librarians in Nigeria: A survey. Health information and libraries journal. Vol.39(2), 155-165

International Federation of Library Association (2020). COVID-19 and the Global Library Field: Statement by the IFLA President and Secretary General. https://www.ifla.org/ last accessed 8/5/2024

Karioja, E. (2013). How to evaluate libraries' sustainability? an approach to an evaluation model and indicators. Available at http://library.ifla.org/114/1/115b-karioja-en.pdf

Lawal-Solarin, Esther.O and Allison, Goodluck O. and Justice, Justice I. (2018) *Contemporary Library Automation Issues In Nigeria.* http://eprints.covenantuniversity.edu.ng/id/eprint/11391. last accessed 8/3/2024

Mapulanga P. (2014). Prospects and constraints of staff development in the University of Malawi Libraries. *The Bottom Line: Managing Library Finances*, 27(1): 29-41.

Obuh, A.O. (2019). Roles and proficiency of Systems librarian as manager of library systems in academic libraries in Nigeria. Journal of library and information science. Vol.10(1) 102-113.

Ogunbodede, K. F., & Cocodia, B. (2023). Promoting the Use of Open Educational Resources by Librarians in University Libraries in Nigeria. *The International Information & Library Review*, 55(4), 329–338. https://doi.org/10.1080/10572317.2023.2183556

Okuonghae, O. and Achugbue, E.I. (2022), 'Digital librarianship practice and open access technology use for sustainable development in Nigeria', *Digital Library Perspectives*, Vol. 38 No. 3, pp. 318-331. https://doi.org/10.1108/DLP-01-2021-0007

Omopupa, K.T., Adedeji, A.A., Kehinde, A.A., Abdulsalam, A.U., Abubakar, H. (2020). Comparative study of Koha usage in Bowen University and University of Ilorin libraries. *J. University Res.* **3**(3), 98–106

Onwubiko, E.C. (2023). Integrated Library System as a Panacea to Sustainable Library Services Delivery in an Era of Pandemic. *Global Perspectives on Sustainable Library Practices.* P:14

Owolabi, K.A., Okorie, N.C., Yemi-Peters, O.E., Oyetola, S.O., Bello, T.O. and Oladokun, B.D. (2022), Readiness of academic librarians towards the use of robotic technologies in Nigerian university libraries, *Library Management*, Vol. 43 No. 3/4, pp. 296-305. https://doi.org/10.1108/LM-11-2021-0104

Strike: SSANU, NASU shut down universities nationwide. (28, October,2024). https://www.vanguardngr.com/2024/10/strike-ssanu-nasu-shut-down-universities-nationwide/.

Tella, A., Okojie, V., Abdullahi, F., & Ajani, Y.A. (2022). The Future of Libraries in Nigeria during the Fourth Industrial Revolution. portal: Libraries and the Academy, 22, 547 - 558.

Tella, A.& Akande, T.O. (2018). ICT knowledge and skills required for recruitment of academic librarians in the digital age. *Library philosophy and practice* (e-journal)

Tella, A., Dina, N., Olaniyi, O.T., Memudu, S.A., Oguntayo, S.A. (2017). Assessment of the use of Koha library software in four selected university libraries in Nigeria. J. Applied Inf. Science Technol. **10**(2), 1–15

Ukwoma, S.C. and Onyebinama, C.O. (2021), 'Challenges and opportunities of facilitating access and use of open access resources to users by librarians in federal and state universities in Nigeria', *Library Management*, Vol. 42 No. 8/9, pp. 481-497. https://doi.org/10.1108/LM-03-2020-0039

Zubair, M.S., Ikwe, J.A., & Edet, G.E. (2023). The Experience of Using Virtua Integrated Library System in Nigerian University Libraries. The Journal of Academic Librarianship.Vol 49(4).

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Digital skills development of university library professionals in Nigeria: testing the UTAUT model

Gbenga Adetunla and Gobinda Chowdhury

DOI: https://doi.org/10.47989/ir30iConf47170

Abstract

Introduction. The UTAUT model has been famous for predicting technology usage. This study tested the UTAUT model on the intention to develop digital skills of university library professionals in Nigeria.

Method. Data was collected through questionnaire from 453 respondents and analysed using descriptive and inferential statistics.

Result. Findings revealed that the university library professionals in Nigeria lack digital skills to carry out some operational tasks, especially serials, acquisition and some cataloguing tasks using z39 and OCLC. This study has shed light on the current digital skills level and factors that impede skills development, such as gender, education, years of experience and career cadre, using the UTAUT model.

Conclusion. This study has upheld the validity, consistency, and expansion of the model beyond technology usage to digital skills development intentions. Recommendations were made to library management and related stakeholders in Nigeria to put to consideration the identified factors as guide-light for planning and decision making on digital capacity development for their staff.

Introduction

The COVID-19 pandemic disrupted university library operations globally, forcing many libraries to shut down in compliance with government restrictions. The International Federation of Library Associations (IFLA) advised libraries to transition to virtual services, emphasizing the importance of digital services to continue operations. A comparative study by Adetunla et al. (2023) highlighted that developed countries especially the UK leveraged technology to maintain operations while many university libraries in Nigeria struggled to function due to the absence of a skilled workforce capable of managing digital tools (Ameh et.al, 2021; Kasa & Yusuf 2022). The success of technology deployment in libraries is heavily dependent on the quality of the workforce, which the current Nigerian digital environment lacks (Beer & Mulder 2020; Ajani et.al., 2023). Nigerian library staff need to develop operational digital skills to effectively utilize library management technologies and improve services such as cataloging and metadata, circulation and patron, institutional repository management etc. (Okuongehe & Achugbue 2020; Corrado, 2023).

The UTAUT (Unified Theory of Acceptance and Use of Technology) metamorphosed from TAM (Technology acceptance model) by Davis (1989). This model predicts that perceived usefulness and perceived ease of use positively affect adoption of technology. However, the model experienced changes over the years. According to Venkatesh et al. (2003) eight different models compete to explain the use of technologies, in adapting all models into one, the UTAUT model emerges as an alternative and was later expanded to UTAUT 2 model (Venkatesh et al., 2012). The UTAUT model identifies four key determinants—performance expectancy, effort expectancy, social influence, and facilitating conditions—that affect an individual's intention to use technology. Performance expectancy refers to the belief that using a system will improve job performance, while effort expectancy reflects how easy the system is to use. Social influence relates to the belief that using technology will enhance one's status in a social group, and facilitating conditions are the resources and infrastructure available to support technology use. Also, moderators such as gender, age, experience, and voluntariness play a vital role in shaping technology adoption (Diaz & Loraas, 2010; Chang, 2012; Keong et al., 2012).

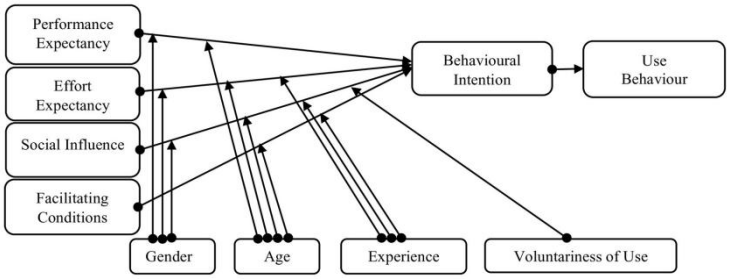


Figure 1. UTAUT model (Venkatesh et al., 2003)

Since the constructs of UTAUT model influence the intention to use technology, this study posits that it should encourage the development of digital skills necessary for technology usage. Research has highlighted that factors such as gender, experience, and IT background are key determinants of the intention to develop digital skills. (Heerwegh et al., 2016; Izuagbe et al., 2019; Couto, 2021; Blut et al., 2022). Arising from on these findings and recognizing that university library professionals in Nigeria lack the digital skills needed to provide services in this digital age, this study aims to explore how the UTAUT model influences the intention to develop digital skills among library professionals. Based on this background, the following research questions and hypotheses were generated:

RQ1. What is the current task-bases digital skills possessed by library professionals in Nigeria.?

RQ2. What are the staff-side factors of the existing digital skill development programs in Nigeria?

H1: there are no gender differences in digital skills level possessed by university library professionals in the use of library automation in Nigeria.

H2: there is no correlation between years of work experience and the level of digital skills possessed by university library professionals in Nigeria for using library automation.

H3: there is no significant correlation between the level of education and the digital skills possessed by university library professionals in using library automation in Nigeria.

H4: there are no significant career cadre differences in the digital skills possessed by university library professionals in using library automation in Nigeria.

Methodology

Research design and sampling technique

This study adopted a survey design using quantitative procedures to collect data through questionnaires. A multi-stage sampling technique was used for this study (see figure 2 below). The first stage involved the use of cluster sampling to put the universities to the 3 geopolitical regions in the South of Nigeria. The second stage involved a random sampling technique to select 3 universities from each region to represent the 3 generations of universities in Nigeria and selection of 3 privately funded universities. and in the third stage, a total enumeration technique was used to sample the 453 professionals (all librarians and library officers) of these 12 universities.

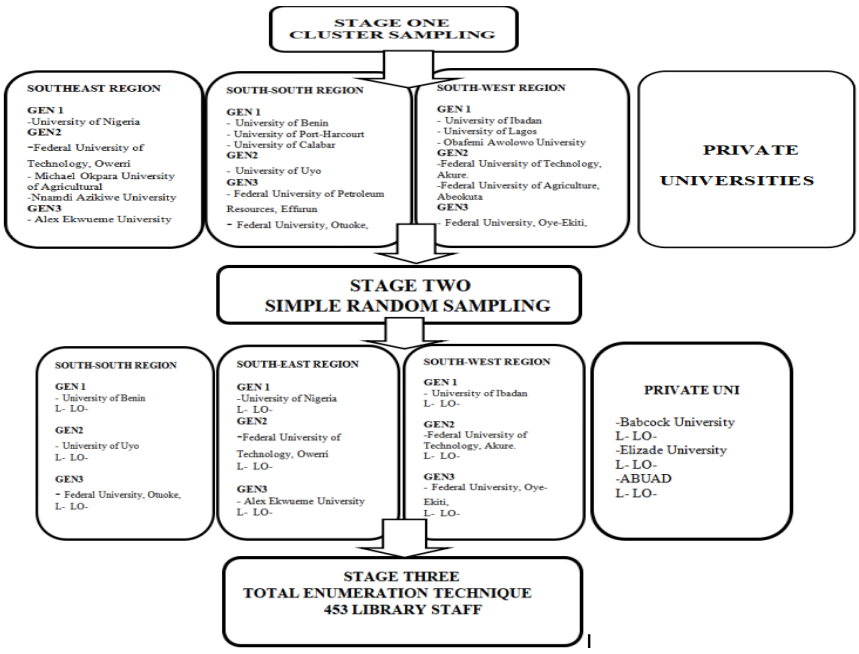


Figure 2. Sampling technique

Quantitative data was collected using structured questionnaires. 453 questionnaires were administered to the respondents across 12 universities in Nigeria; 432 questionnaires were retrieved, and 408 questionnaires were properly filled and found useful for the analysis of this study. Data collected for the study was analyzed using SPSS version 23. T-test of independent sample and Pearson product moment correlation (PPMC) was used to test the hypotheses of the study. The hypotheses were set at 0.05 level of significance.

Presentation of result and discussion

Demographics distribution of respondents

Institution	**FUTA**	**FUTO**	**UNN**	**FU'OYE**	**FU'OTU**	**AEU**	**UNI'BEN**	**UI**	**UNI'UYO**	**BABCOCK**	**ELIZADE**	**ABUAD**	**TOTAL**
Frequency	17	80	78	10	18	22	32	70	30	25	10	16	408
Percentage	4.2	19.6	19.1	2.5	4.4	5.4	7.8	17.2	7.4	6.1	2.5	3.9	100%
Gender	**male**		**female**		**prefer not to say**								
Frequency	168		226		14								408
Percentage	41.2		55.4		3.4								100%
Designation	**Librarian**			**Library Officer**									
Frequency	164			244									408
Percentage	40.2			59.8									100%
Education	**Diploma**			**Bachelors**			**Masters**			**PhD.**			
Frequency	159			156			65			28			408
Percentage	38.9			39.0			15.9			6.9			100%
Work Experience	**0 – 5**			**6 – 10**			**11 above**						
Frequency	83			202			123						408
Percentage	20.3			39.5			40.1						100%

Table 1. Analysis of demographics distribution of respondents

Table 1 shows the demographic analysis of the respondents. Based on institutions, FUTO has the most respondents while Elizade University has the least respondents. Result based on gender shows that there is more female representation. Analysis of designation shows that library officers are more than librarians. Based on their educational qualifications, it was shown that staff that hold Diploma are more, and PhD comes as the least. Analysis of work experience shows that most of the respondents had between 6 to 10 years' working experience.

Research questions

Research question 1: what are the task-based digital skills possessed by university library professionals in Nigeria?

OPERATIONAL TASK DIGITAL SKILLS						
Acquisition operation	Manage vendors	Manage budget	Create purchase order	Manage order/ receipt	Handle purchase suggestions	
$\bar{X}$	1.79	1.84	1.82	1.79	1.71	
SD	.578	.607	.719	.661	.746	
Remark	unskilled	unskilled	unskilled	unskilled	unskilled	
Cataloging	Edit bibliographic records	Manage OPAC	Generate call mark/ accessioni ng	General cataloging	OCLC importing	Z39.50 tool
$\bar{X}$	3.50	3.48	3.45	3.43	1.83	1.75
SD	1.337	1.345	1401	1.418	.684	.728
Remark	skilled	skilled	skilled	skilled	unskilled	unskilled
Serial	Manage subscription	Create new issues/Manage collection	Check subscription expiration	Manage numbering pattern	Claim late issues & manage frequency	
$\bar{X}$	3.43	1.87	3.50	1.91	1.82	
SD	1.416	.650	.1.343	.661	.656	
Remark	skilled	unskilled	skilled	unskilled	unskilled	
Circulation	User registration	Charge & discharge	Renew & prompt due dates	Inter-library operations	Generate reports	Hardware usage
$\bar{X}$	3.37	3.46	3.43	3.47	3.46	3.46
---SD	1.439	1.362	1.400	1.355	1.402	1.366
Remark	skilled	skilled	skilled	skilled	skilled	skilled
Patron	Create new user category	Register new patron profile	Manage/ed it patron account	Staff/patron permission		
$\bar{X}$	3.42	3.40	3.43	3.42		
SD	1.408	1.423	1.393	1.391		
Remark	skilled	skilled	skilled	skilled		

Table 2. Analysis of task-based digital skills possessed by university library professionals in Nigeria

Table 2 shows university library professionals in Nigeria lack digital skills for acquisition and serials tasks. While they possess some skills for cataloging tasks, it is evident they are not skilled in using the Z39.50 tool and OCLC. However, they are skilled in the circulation and patron tasks. The extent of automation in the university libraries in Nigeria shows the level of contact the professionals have with the various operations of their library management systems hence the reason for an average level of skill in cataloging and circulation routine and a low level of skill in the other library routine they are less exposed to such as acquisition and serial routines (Ajani & Buraimo, 2021).

Research question 2: what are the staff-side factors that impede the existing digital skill development programs for university library professionals in Nigeria

Challenges of Digital Skills	N	$\bar{X}$	SD	Remark
Personal factors				
I have phobia for computers and technology	408	1.40	.583	Disagree
My age affects the use of computers and technology	408	1.66	.532	Disagree
My religious belief is averse to the use of computers	408	1.70	.561	Disagree
I have health challenges that requires me to avoid technology	408	1.70	.568	Disagree
I am not financially buoyant to afford training	408	3.86	.621	Agree
My formal LIS training excluded training on digital technology	408	3.65	.624	Agree
Gender related factors				
Family commitments give me little time for training	408	3.67	.649	Agree
My marriage reduces my digital skills training opportunity	408	3.65	.670	Agree
My gender provides little opportunity for digital skills training	408	3.41	1.422	Agree
Organizational factors				
My rank in the office reduces my training opportunities	408	1.66	.672	Disagree
I am rarely promoted and there is no need to attend any digital skills training	408	1.39	.581	Disagree
My library management does not encourage digital skills training	408	3.67	.649	Agree

Table 3. Staff-side factors that impede the existing digital skill development programs for university library professionals in Nigeria

Personal factors: library staff reported that their prior formal library school education lacked courses on digital library technologies. This curriculum gap is a barrier to developing digital skills. Many library schools still offer outdated courses, and even where digital courses exist, they tend to be theoretical due to a lack of practical facilities and qualified instructors. Ikolo and Nongo (2022) assert that same is still the case. Also, staff members often cannot afford self-sponsored digital training, particularly when international training opportunities require travel and foreign currency expenses, making it difficult for those from developing countries like Nigeria to access these opportunities. Aregbesola et.al, (2019) alluded that the challenge of affordability caused by foreign exchange subsumes most developing countries. Additionally, gender-related challenges, particularly for women in developing countries especially Africa, affect their ability to develop digital skills. Female professionals in Nigeria are faced with care-giving responsibilities (e.g., childbirth and family care), which often lead to career interruptions and limit their opportunities for digital skills development compared to their male counterparts (Adjah & Walt, 2019; Akhtar & Soroya, 2021).

Organizational factors: many library staff perceive a lack of commitment from university library management to support digital skills development. There is limited financial backing for training programs, with staff development often not prioritized. Akintola et.al. (2022) alluded that library personnel in the universities perceived that library management was not committed to capacity building as there is irregular sponsorship for training from their university. Due to economic challenges facing the universities and low budgetary allocations for staff development, university libraries struggle to sponsor training programs for their staff. Additionally, the non-payment of salaries affects staff morale and further discourages participation in training opportunities. Mapulanga (2014) observed that the dwindling economy in most developing countries and especially in Nigeria has put the universities management in financial quagmire of meeting their obligations on training and wages payment (Vanguard Newpaper, 2024). Even when training

opportunities come, non-payment of salary affects their morale and motivation to attend training programs.

Hypotheses

Hypothesis 1: there are no significant gender differences in digital skills level possessed by university library professionals in Nigeria.

Gender	**N**	**Mean**	**SD**	**Df**	**$t_{(cal)}$**	**$t_{(tab)}$**	**Decision**
Male	168	3.72	1.321	392	4.32	1.96	S
Female	226	3.43	1.285				
P<0.05 level of significance					*S = Significant*		

Table 4. T-test analysis of gender difference in digital skills level possessed by university library staff in the use of library automation in Nigeria

Table 4 revealed that there was significant gender difference in digital skills level possessed by library staff in Nigeria. The result shows that male professionals have more digital skills than the women professionals. This outcome aligns with the findings of Deyrup (2014) and Ashiq and Warraich (2023) who identified that one of the challenges of female professionals in the libraries is lack of digital skills and the digital space is gradually moving towards masculinity. Golub and Jose (2009) analyzed the professional trends and reveals a new pattern of dual career tracking in which men are associated with technology and women with management.

Hypothesis 2: there is no significant correlation between the years of work experience and digital skills level possessed by library staff in the use of library automation in Nigeria.

Correlations			
		Work Experience	Digital Skills Level
Work Experience	Pearson Correlation	1	.050
	Sig. (2-tailed)		.310
	Sum of Squares and Cross-products	202.078	7.392
	Covariance	.497	.018
	N	408	408
Digital Skills Level	Pearson Correlation	.050	1
	Sig. (2-tailed)	.310	
	Sum of Squares and Cross-products	7.392	106.586
	Covariance	.018	.262
	N	408	408

T-test analysis of difference in the years of experience between 0 to 10 years and 11 years and above

Years of Experience	**N**	**Mean**	**SD**	**df**	**$t_{(cal)}$**	**$t_{(tab)}$**	**Decision**
0 – 10	285	3.75	1.497	406	2.87	1.96	S
11 years and above	123	3.40	1.172				

P<0.05 level of significance *S = Significant*

Table 5. Analysis of correlation between the years of work experience and digital skills level possessed by library staff in the use of library automation in Nigeria

Table 5 revealed a significant positive correlation between years of work experience and the digital skills level of library staff in Nigeria. Further t-test analysis revealed that library staff with 0 to 10 years of experience possess higher digital skills than those with 11 years and above. This is significant because most of the study's participants fall into the 0 to 10 years category, suggesting that many of them are younger, possibly digital natives, who interact more frequently with technology in their daily work. In contrast, the group with over 10 years of experience primarily represents the library leadership, responsible for decision-making. The study highlights that the library leadership in Nigerian universities may lack the necessary digital skills due to outdated prior library school training which were not aligned with the demands of the current digital environment. This has left the leadership underprepared to address the challenges posed by new technologies (Ariole et al., 2017; Olubiyo, 2022).

Hypothesis 3: there is no significant correlation between the level of education and digital skills level possessed by library staff in Nigeria.

Hypothesis 4: there is no significant difference in the career cadre on the digital skills level possessed by library staff in Nigeria.

Correlations			
		Level of Education	Digital Skills Level
Level of Education	Pearson Correlation	1	.055
	Sig. (2-tailed)		.268
	Sum of Squares and Cross-products	221.490	11.863
	Covariance	.544	.029
	N	408	408
Digital Skills Level	Pearson Correlation	.055	1
	Sig. (2-tailed)	.268	
	Sum of Squares and Cross-products	11.863	210.578
	Covariance	.029	.517
	N	408	408

T-test analysis of difference in the career cadre

Career Cadre	**N**	**Mean**	**SD**	**df**	**$t_{(cal)}$**	**$t_{(tab)}$**	**Decision**
Librarian	12	2.25	.452	406	3.029	1.96	S
Library Officer	396	1.73	.592				

P<0.05 level of significance *S = Significant*

Table 6. Analysis of correlation between the level of education and digital skills level possessed and t-test analysis of difference in the career cadre and the digital skills level possessed by library staff in Nigeria

Table 6 indicates a significant positive correlation between the level of education and the digital skills level of library staff in Nigeria. This suggests that higher levels of formal education contribute to better digital skills. The study also distinguishes between two career cadres: librarians and library officers and reveals a significant difference in digital skills between these two groups, with librarians possessing higher levels of digital skills compared to library officers. Librarians typically hold at least a bachelor's degree in library and information science or a related field, while library officers have a diploma in the same field. These findings align with Inyang and Mngutayo (2018) and Ibegbulam and Eze (2016), who noted that formal education is a key avenue for acquiring the required skills and so, paraprofessionals, like library officers, tend to have lower digital skills.

Connecting the findings to the UTAUT

Figure 3 shows a self-developed adaptation of the UTAUT model based on the findings of this study.

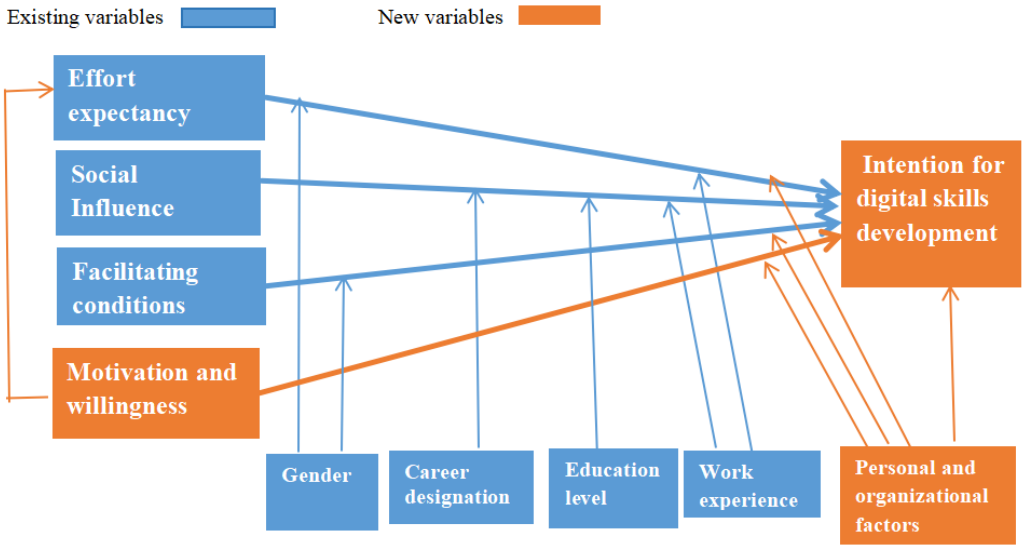


Figure 3. Shows a self-developed adaptation of the UTAUT model

Effort expectancy: this concept refers to the ease with which library professionals can use technology operational tasks. To improve the ability to use technology with ease, digital skills development is essential. Gender, work experience and organizational factors were found to be significant moderating factors. Female staff and those with more years of experience exhibiting lower digital skill will hinder task. Also, lack of motivation and incentives from the organization influence willingness to develop skills which is crucial for operational use of technology.

Social influence: social influence pertains to how career position, education, and social status can affect technology use. The study highlights that staff with higher education levels, particularly senior staff (librarians with master's or PhD qualifications), are more skilled in using technology than those with lower educational qualifications. Career cadre and education level were identified as moderators of social influence, impacting the intention to develop digital skills.

Facilitating conditions: conditions such as technical infrastructure, organizational support, and personal factors (e.g., financial stability, health, willingness to learn) are essential to overcoming barriers to digital skills development. Gender-related challenges (e.g., care-giving responsibilities) and organizational factors (e.g., incentives, remuneration) also play a significant role in influencing staff's intention to develop digital skills.

Motivation and willingness: this new emergence highlights the importance of addressing personal and organizational concerns to increase staff motivation and willingness to improve their digital skills. Key factors like financial stability, timely payment of wages, workplace gender-related issues, and promotion opportunities play a significant influence on digital skills development.

Conclusion and recommendations

The study demonstrates that the UTAUT model remains relevant in understanding technology adoption, especially in the context of digital skills development within the library and information field. This study has tested the UTAUT in a different but similar terrain by extending the discussion beyond technology usage to digital skills development. This study has upheld the validity, consistency, and expansion of the model, introducing new moderating factors such as personal and organizational commitments, as well as variables such as motivation and willingness, which enhance digital skills development. Beyond theory, this study has shed light on the gaps and barriers to digital skills development among Nigerian university library professionals, addressing

these gaps will enable university libraries in Nigeria to provide up-to-date digital services and ultimately improving the quality of education and research output in Nigeria. Following recommendations are made:

Policy and planning: library administrators and relevant stakeholders should consider the identified factors impeding digital skills development to guide policy formation, planning, and decision-making.

Institutional action: university library administrators should conduct a detailed analysis of the digital skills needs within their institutions and implement tailored training methods to address the specific skill gaps among staff.

Gender equity in training: there should be a focus on providing equitable digital skills development opportunities for female professionals, addressing gender-related barriers to advancement.

Staff motivation: university management should prioritize staff motivation, ensuring that factors such as financial incentives, career progression, and recognition are in place to encourage staff to develop the digital skills necessary for enhancing their performance and the overall output of the institution.

About the authors

Gbenga Adetunla is a Doctoral researcher at the University of Strathclyde, UK. His research focuses on sustainable digital library operations. He can be contacted at gbenga.adetunla@strath.ac.uk

Gobinda Chowdhury is a Professor in Information Science at University of Strathclyde, UK. His research focuses on digital libraries and information services. He can be contacted at gobinda.chowdhury@strath.ac.uk

References

Adjah, A.O. &Walt, V.T. (2019). Career progression of female librarians in public university libraries in Ghana. *Journal of Librarianship and Information Science*, 51, 331 - 345.

Adetunla, G., Rasmussen Pennington, D., Chowdhury, G. (2023). Automation of University Library Operations: An Analysis of the COVID-19 Pandemic Experience in the United Kingdom and Nigeria. *LNCS* 13972, pp. 1–9,

Ajani, Y.A., Dunmade, A.O., Tella, A., & Adeniran, C.O. (2023). Information Professionals of the Future and their Prospects in the Era of Fourth Industrial Revolution: The Need for Transformative Potential in Nigeria. *Mousaion: South African Journal of Information Studies*.

Ajani, F.O. and Buraimo, O. (2021). Perceived impact of automation on university library services by library personnel in Southwest, Nigeria. Information Development. Vol.38(2), 179-191

Akhtar, G & Soroya, M. S. (2021). Factors Influencing Career Routes of Female Librarianship: A Literature Review. *Library Philosophy and Pra*

Akintola, B.O., Adetomiwa, B., & Onifade, F.N. (2022). Staff Training and Career Development among Library Staff in Federal Universities in Southwest, Nigeria. *International Information & Library Review*, 54, 315 - 324.

Ameh, G J, Ukwuoma, H C And Oye, P O. (2021). Covid-19 Pandemic and Evolving Library and information Services: Lessons for Nigeria. International Journal of Development and Management Review Vol. 16, No. 1.

Aregbesola, A., Van der Walt, T, Owolabi, S., Idiegbeyan-ose, J., Okocha F. and Eyiolorunshe, T. (2019). Infopreneurship in a developing country: Opportunities and Challenges. 1st International Conference on Sustainable Infrastructural Development 24–28 June 2019, Covenant University, Canaan Land, Ota, Nigeria. *IOP Conference Series: Materials Science and Engineering*, Vol 640, DOI 10.1088/1757-899X/640/1/012123

Ariole, I.A., Benson, O.V. & Okorafor, K. (2017). Expectations of Library Schools in the Preparation of Future Library Environment: Perspectives of African Countries Information and Knowledge Management Vol.7 (12)

Ashiq, M and Warraich, N.F. (2023) Challenges and strategies to attain leadership positions for female library professionals: A narrative review. *The Journal of Academic Librarianship*, Vol 49(4),

Beer P and Mulder RH. (2020) The Effects of Technological Developments on Work and Their Implications for Continuous Vocational Education and Training: A Systematic Review. Front. Psychol. 11:918.

Blut, M., Chong, A., Tsiga, Z., & Venkatesh, V. (2022). Meta-Analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): Challenging its Validity and Charting a Research Agenda in the Red Ocean. *Journal of the Association for Information Systems*, 23(1), 13-95.

Chang, A. (2012) Utaut and Utaut 2: A Review And Agenda For Future Research.Journal The Winners, Vol. 13(2)

Couto (2021) Developing digital skills intention: A test of a UTAUT variant model. Masters dessartation, University Institute of Lisbon. Portugal

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340

Deyrup A. (2014). Academic library leadership, second wave feminism, and twenty-first century humanism

Diaz, M. C., & Loraas, T. (2010). Learning new uses of technology while on an audit engagement: Contextualizing general models to advance pragmatic understanding. International Journal of Accounting Information Systems, 11, 61–77. doi: 10.1016/j.accinf.2009.05.001

Golub, E.M., & José, S. (2014). Gender divide in librarianship: past, present, and future. *Library Student Journal*

Heerwegh D. et al., (2016). Exploring the self-reported ICT skill levels of undergraduate science students. Journal of Information Technology Education

Ibegbulam, I. and Eze, J.U. (2016), 'Training needs of paraprofessional library staff in university libraries in South-East Nigeria', *Library Management*, Vol. 37 No. 8/9, pp. 482-495.

Ifijeh, G & Yusuf, F (2020). Covid–19 pandemics and the future of Nigeria's university system: The quest for libraries' relevance. The Journal of Academic Librarianship, 46 (6).

IFLA. (2020) COVID-19 and the Global Library Field: Statement by the IFLA President and Secretary General. Available at https://www.ifla.org/ last accessed 8/5/2024

Ikolo and Nongo, (2022). Training needs of medical librarians in Nigeria: A survey. Health information and libraries journal. Vol.39(2), 155-165

Inyang, E. and Mngutayo, J (2018). Required skills and competences of librarians for effective software application and use in contemporary libraries in Nigeria. *Library philosophy and practice(e-journal)*

Izuagbe et.al (2019) Effect of perceived ease of use on librarians' e-skills: Basis for library technology acceptance intention. *Library & Information Science Research.* Vol 41(3)

Keong, M. L., Ramayah, T., Kurnia, S., & Chiun, L. M. (2012). Explaining intention to use anenterprise resource planning (ERP) system: an extension of the UTAUT model. BusinessStrategy Series, 13(4), 173 - 180. doi: 10.1108/17515631211246249

Kasa, M. G. and Yusuf, A,. (2020). Expof Deyrup An Academic Library During the Covid-19 Pandemic. *Library Philosophy and Practice (e-journal).*4456.

Olubiyo, P. (2022). *Library Education in Africa: A Foundational Perspective*. pp.252-262 DOI:10.4018/978-1-6684-5964-5.ch019

Venkatesh, V. et.al. (2003). *Unified Theory of Acceptance and Use of Technology* (UTAUT). APA PsycTests. Available at https://doi.org/10.1037/t57185-000

Venkatesh, V., Thong, J.Y.L. and Xu, X. (2012) Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, 36, 157-178

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Aprendizaje-Servicio como instrumento para la accesibilidad y la inclusión: resultado de una propuesta colaborativa entre universidades y bibliotecas (*Service-learning as a tool for accessibility and inclusion: results of a collaborative approach between universities and libraries*)

Elena Loreto Olmedo-Pagés y Rosario Arquero-Avilés

DOI: https://doi.org/10.47989/ir30iConf47167

Abstract

Introduction. Teaching innovation projects based on service-learning (SL) have proven to be beneficial in the creation of alliances between organisations committed to pooling their efforts and resources to have a positive impact on their environment.

Implementation proposal. In the framework of a previous project, this article presents a service-learning (SL) project developed between library and information science students and libraries with the aim of adapting materials to Easy to Read in order to improve accessibility for people with hearing disabilities. The project is aligned with the Sustainable Development Goals (SDGs) and aims to raise awareness of the importance of accessibility in library services. Through collaboration, students gain key skills in creating accessible content, while libraries strengthen their inclusive role by providing resources that facilitate equitable access to information.

Conclusions. A sustainable implementation model is proposed that can be replicated by other universities and libraries interested in promoting inclusion and equal opportunities in access to culture and knowledge.

Introducción

Los Estados miembros de las Naciones Unidas establecieron en 2015 un conjunto de 17 metas globales como parte central de la 'Agenda 2030 para el Desarrollo Sostenible', (United Nations, 2015). Los conocidos como Objetivos de Desarrollo Sostenible (ODS) tienen como propósito afrontar los retos y desafíos más urgentes del mundo, abordando dimensiones medioambientales, políticas, económicas y sociales (Naciones Unidas, 2024).

Disponer de estos ODS como marco global para mejorar la sostenibilidad y la equidad en diversas áreas de la sociedad, incluidas la educación y la inclusión social, es de gran relevancia para encauzar la actividad de instituciones educativas y culturales, como pueden ser las bibliotecas. Estas, como agentes de cambio social, juegan un papel decisivo en la eliminación de las dificultades que existen en el acceso a la información para grupos vulnerables, como lo son las personas con discapacidad auditiva (UNESCO, 1994). Las bibliotecas, más que centros de información, son espacios de inclusión, acceso equitativo y alfabetización digital. Para cumplir con ese papel, los profesionales que trabajan en ella deben estar formados en materia de accesibilidad, de modo que puedan dar respuesta a todas las necesidades que puedan presentar sus usuarios.

Una herramienta que puede ser de gran utilidad a la hora de formar a esos futuros bibliotecarios en las aulas de la universidad y a los profesionales que ya ocupan sus puestos en las bibliotecas es la propuesta de proyectos de Aprendizaje-Servicio (ApS). Este método permite que el alumnado adquiera las competencias y conocimientos propios de la materia desde una perspectiva real, a través de la participación activa en la comunidad de la que forman parte por medio de la oferta de un servicio de calidad que contribuya a la mejora y transformación de dicho entorno (Litchke et al., 2020).

Este trabajo pretende visibilizar el impacto que un proyecto de ApS puede tener en el entorno social al identificar necesidades por medio de la colaboración entre diferentes agentes. Aunando los esfuerzos de estudiantes universitarios, personal docente e investigador y profesionales de bibliotecas y unidades de información y documentación, se ha podido detectar la falta de adaptación de la información del entorno bibliotecario para personas con discapacidad auditiva. Con ello, este método se presenta como una importante herramienta para aportar soluciones a los problemas sociales.

Aprendizaje-servicio en la universidad: estrategias para impulsar la accesibilidad en las bibliotecas

El ApS es una metodología pedagógica que combina el aprendizaje de contenidos y competencias con la prestación de servicios a la comunidad con la que trabaja. Se basa en el establecimiento de alianzas con entidades de su entorno para mejorar la realidad, las condiciones de vida de las personas, su inclusión y su participación activa en la sociedad. Esta práctica está estrechamente vinculada a los ODS porque comparten con ellos finalidades que garantizan la paz y el progreso de las sociedades más vulnerables, ya que 've y considera a la sociedad como una red de relaciones interpersonales y de responsabilidad [donde] todos y todas debemos procurar el bien para todos [...]' (Uruñuela, 2018, p. 43). En términos generales, las iniciativas de ApS contribuyen a la sostenibilidad educativa y curricular en la universidad, al tiempo que promueven en los estudiantes un mayor entendimiento, sensibilización, compromiso y avance gradual en la consecución de los ODS a través de las acciones previstas (González et al., 2022).

Entre las características que definen la metodología ApS, destaca la reciprocidad en la mejora entre en aprendizaje y el servicio, ambas se benefician de la otra al crearse un proyecto educativo de utilidad social, que promueve los valores cívicos y éticos y sirve como modelo socioeducativo para la promoción de actividades solidarias (Mendía, 2012). Promueve la formación, la investigación y la sensibilización de la comunidad universitaria al crear un vínculo entre esta y aquellos que reciben el servicio, muchas veces conociendo realidades muy diferentes a las que viven en el aula y

teniendo que adaptarse a necesidades particulares de ese entorno (Folguerias et al., 2013). Es por ello que esta metodología se plantea como la idónea para formar al alumnado en accesibilidad y desarrollar proyectos que promuevan la inclusión de las personas con discapacidad (Jiménez, 2020; Mendía, 2016; Ríos, 2012).

La aplicación del ApS en la Educación Superior, aunque es relativamente reciente, adquirió mucha fuerza desde sus inicios, dando como resultado el compromiso del alumnado con su entorno social inmediato a través de proyectos donde ellos mismos gestionan la respuesta que deben dar a las necesidades de la realidad social a la que se enfrentan (Sotelino et al., 2016). La sensibilización de los estudiantes a través de su participación es otra de las cuestiones en la que coinciden numerosos autores (Jiménez, 2020; Puig, 2019; Mayor y López, 2017). Por tanto, se puede afirmar que el ApS es una metodología efectiva y muy conveniente para el establecimiento de alianzas destinadas a la promoción de la accesibilidad y la inclusión en el entorno educativo y cultural.

Con esta premisa de base, se elaboró un proyecto de ApS auspiciado y financiado por la Universidad Complutense de Madrid, titulado 'Aprendizaje Servicio en el área de Biblioteconomía y Documentación: un proyecto generador de proyectos en bibliotecas, archivos y centros de documentación' (Arquero et al., 2020; Arquero, 2024). Este proyecto se planteó como experiencia docente dentro del marco de la asignatura obligatoria 'Dirección y gestión de proyectos en unidades de información y documentación' del Máster en Gestión de la Documentación, Bibliotecas y Archivos de la Facultad de Ciencias de la Documentación de la Universidad Complutense. Personal docente e investigador de la mencionada universidad y profesionales de bibliotecas y unidades de información y documentación dependientes de diversas administraciones (Ministerio de Educación, Cultura y Deporte y Subdirección General del Libro de la Comunidad de Madrid) participaron en este proyecto, a lo largo de las diferentes ediciones del proyecto (desde su primera edición en el curso 2019-2020 hasta el curso académico actual 2024-25), con el objetivo de ofrecer a los estudiantes un punto de partida a partir del cual proponer proyectos relacionados con la asignatura y los ODS (Arquero et al., 2020).

Gracias a la realización de este proyecto, el alumnado de la mencionada titulación recibió formación sobre los ODS y el diseño y gestión de proyectos, asistió a visitas programadas con los profesionales de diferentes bibliotecas y desarrolló proyectos basados en las necesidades detectadas en estos entornos y en los ODS que buscaran mejorar en algún aspecto las condiciones de la biblioteca en cuestión (Arquero, 2024; Arquero et al., 2020). Es así como, por medio del trabajo colaborativo, se identificó un campo en el que las bibliotecas todavía deben trabajar para garantizar la accesibilidad a la información que contienen.

Accesibilidad para personas con discapacidad auditiva: respuesta a las necesidades por medio de la Lectura Fácil

Las bibliotecas, al igual que las universidades, contribuyen a la consecución del ODS 4 de Educación de calidad por medio de la inclusión educativa, proporcionando recursos educativos a todos los niveles y promoviendo el desarrollo de habilidades de forma equitativa. Las bibliotecas deben disponer de materiales y servicios que satisfagan las necesidades informativas y formativas de todos sus usuarios (IFLA, 2010; IFLA y UNESCO, 2001). Al mismo tiempo, responden al ODS 10 de Reducción de las desigualdades al procurar un acceso equitativo a la información y los servicios educativos y de ocio. Es numerosa la normativa que se ha establecido a nivel internacional y nacional sobre la accesibilidad física y web a los contenidos que ponen a disposición las bibliotecas. Estas normas promueven las medidas necesarias para la eliminación de las barreras a las que se enfrentan las personas con discapacidad (Martínez, 2007).

La comunicación con la comunidad sorda ha sido uno de los principales problemas afrontados desde las bibliotecas, y es por ello por lo que existen pautas específicas para dar servicio a usuarios con discapacidad auditiva. Los Lineamientos para los Servicios Bibliotecarios para Personas Sordas

(2000) son un claro ejemplo de ello, un documento de alcance internacional que sirve como guía en el desarrollo de lineamientos nacionales para servicios bibliotecarios dirigidos a personas con discapacidad auditiva (Day, 2000). A nivel nacional, en el caso de España, estas directrices tienen su reflejo en documentos como la Guía para hacer bibliotecas accesibles (2018), que tiene un apartado específico para cada tipo de discapacidad y que hace especial hincapié en que toda la información que ofrece la biblioteca ya sea en físico o a través de una web, sea de fácil comprensión. Además, hace mención a la disponibilidad libros en Lectura Fácil, como medio de accesibilidad que beneficia a personas con discapacidad pero, también, a otros usuarios que, por ejemplo, pueden no conocer completamente el idioma (Calícrates, 2018). Específicamente diseñadas para las personas con discapacidad auditiva, las Pautas para bibliotecas españolas al servicio de usuarios/as con diversidad funcional auditiva signantes (2016) ofrece como solución a las dificultades en la lectura de este colectivo una colección de libros en Lectura Fácil (Manzanares, 2016).

Muchas bibliotecas han puesto a disposición de sus usuarios una colección de libros en Lectura Fácil, en formato papel y digital, por medio del préstamo a nivel individual o para entidades que organizan clubs de lectura (Down España, 2017; Casals, 2006). Sin embargo, gracias a la implantación del proyecto Aprendizaje Servicio en el área de Biblioteconomía y Documentación: un proyecto generador de proyectos en bibliotecas, archivos y centros de documentación, se pudo identificar un campo que debe ser trabajado para mejorar la accesibilidad de las personas con discapacidad auditiva en las bibliotecas. La necesidad detectada en este caso fue que, a pesar de disponer de colección de Lectura Fácil, el resto de los productos de la biblioteca que garantizan la autonomía de estas personas en sus instalaciones no tienen una versión en Lectura Fácil. La señalética, la Carta de Servicios y otros materiales que la biblioteca crea para informar a sus usuarios no es accesible para este colectivo. Por este motivo, gracias al planteamiento de este proyecto como generador de proyectos, se realiza ahora una propuesta de Aprendizaje Servicio en la que se resuelva esta necesidad por medio de la colaboración entre diferentes entidades.

Propuesta de implementación: un proyecto para la inclusión de las personas con discapacidad auditiva en las bibliotecas por medio de la Lectura Fácil

A continuación, se presenta una propuesta de implementación de un estudio de caso que se contextualiza en el proyecto marco citado de Aprendizaje Servicio en el área de Biblioteconomía y Documentación: un proyecto generador de proyectos en bibliotecas, archivos y centros de documentación (Arquero et al., 2020; Arquero, 2024), cuya continuidad se está abordando en el curso académico 2024-25. Se trata de una propuesta de proyecto derivada de la detección de necesidades en bibliotecas para que instituciones educativas y culturales puedan adaptarlo a sus características propias y llevarlo a cabo.

El objetivo principal es el establecimiento de una alianza entre los estudiantes universitarios de la disciplina de Biblioteconomía y Documentación y las bibliotecas para adaptar materiales informativos a Lectura Fácil, promoviendo así la accesibilidad de personas con discapacidad auditiva. A través de este proyecto, se busca mejorar el acceso a la información y fomentar la inclusión social en las bibliotecas, mientras los estudiantes desarrollan habilidades prácticas, aumentan su conciencia social y contribuyen al cumplimiento de los ODS.

Metodología del proyecto

Para llevar a cabo este proyecto, en primer lugar, deberán identificarse los centros que van a colaborar para su implementación y concretar las responsabilidades de cada parte. Los agentes participantes se agruparán en los siguientes bloques: por un lado, estará el alumnado con el que se pretende trabajar, por otra parte, la biblioteca que se beneficiará de los nuevos materiales y,

finalmente, otros centros interesados que puedan aportar su conocimiento en materia de Lectura Fácil.

Posteriormente, se procederá a la implantación propiamente dicha del proyecto de Aprendizaje Servicio, siguiendo cada una de las fases que se plantean en el siguiente apartado.

Por último, se llevará a cabo una evaluación detallada de los resultados de la implantación de este proyecto desde una perspectiva bilateral. Por una parte, se evaluarán los resultados obtenidos del trabajo del alumnado por medio de metodologías cuantitativas y cualitativas que permitan calificar el aprendizaje y el impacto del proyecto. Por otra, se medirá la repercusión de la disponibilidad del nuevo material adaptado a Lectura Fácil en la biblioteca en cuestión, así como la mejora del servicio que los profesionales pueden ofrecer a sus usuarios.

Estructura del proyecto ApS

Este proyecto se plantea en una serie de fases.

Fase 1: planificación y formación

1.1. Identificación de necesidades: reunión entre la universidad y las bibliotecas participantes para identificar las necesidades específicas en cuanto a materiales que deban ser adaptados a Lectura Fácil (guías de usuario, normativas, folletos, etc.).

1.2. Formación de los estudiantes en Lectura Fácil: los estudiantes recibirán talleres o cursos introductorios sobre el concepto de Lectura Fácil, metodologías de adaptación de textos y normativas de accesibilidad. Este módulo será impartido por expertos en accesibilidad o profesionales en bibliotecas especializadas.

1.3. Definición de roles y responsabilidades: se formarán equipos de estudiantes que trabajarán bajo la tutoría de un profesor y un bibliotecario. Cada equipo será responsable de un conjunto de materiales a adaptar.

Fase 2: desarrollo del proyecto

2.1. Selección y revisión de materiales: los estudiantes seleccionarán y revisarán los materiales proporcionados por las bibliotecas, como guías de usuario, normativas de acceso o materiales de divulgación, identificando las partes que necesitan adaptación a Lectura Fácil.

2.2. Proceso de adaptación: los estudiantes adaptarán los materiales siguiendo las pautas de Lectura Fácil que hayan aprendido en los primeros talleres.

2.3. Validación con usuarios finales: una vez adaptados los materiales, se realizarán pruebas con usuarios de las bibliotecas que tengan discapacidad cognitiva, auditiva o dificultades de lectura. Estas pruebas permitirán verificar si los materiales cumplen con los criterios de Lectura Fácil. La retroalimentación recibida será utilizada para mejorar los textos resultantes.

Fase 3: implementación y evaluación

3.1. Publicación y distribución de los materiales: los materiales adaptados se publicarán en formato físico o digital (sitio web de la biblioteca o repositorios digitales accesibles), asegurando su disponibilidad para los usuarios de las bibliotecas.

3.2. Formación al personal bibliotecario: los estudiantes también organizarán talleres de sensibilización para el personal bibliotecario sobre la importancia de la Lectura Fácil y cómo dar una mejor respuesta a los usuarios que necesiten estos recursos.

3.3. Evaluación del impacto: se llevará a cabo una evaluación final donde se medirá el impacto del proyecto en términos de:

- Número de materiales adaptados.

- Cantidad de usuarios que han utilizado los materiales en Lectura Fácil.
- Retroalimentación de usuarios y bibliotecarios sobre la utilidad y accesibilidad de los materiales.
- Reflexión por parte de los estudiantes sobre las habilidades aprendidas y el impacto social del proyecto.

El modelo planteado es flexible, es decir, puede ajustarse a las características y recursos de cada universidad o biblioteca que lo implemente.

Conclusiones

La implementación de un proyecto de ApS entre profesionales bibliotecarios, personal docente e investigador y estudiantes universitarios de Biblioteconomía y Documentación para la creación de materiales adaptados a Lectura Fácil representa una iniciativa transformadora tanto para los estudiantes como para la comunidad. Este tipo de colaboración no solo enriquece la formación académica de los estudiantes, al permitirles aplicar conocimientos teóricos en un contexto real y de impacto social, sino que también contribuye a hacer que las bibliotecas sean más accesibles para personas con discapacidad auditiva y, con ello, más inclusivas. Estas medidas de accesibilidad pueden, además, beneficiar al mismo tiempo a otros usuarios, como personas con discapacidad cognitiva o con dificultades de comprensión lectora.

Al fomentar el acceso equitativo a la información, este proyecto avanza hacia el cumplimiento de los ODS, en especial los relacionados con la educación inclusiva y la reducción de desigualdades.

El éxito de este proyecto radica en su enfoque práctico y colaborativo, que no solo responde a una necesidad social urgente, sino que también fortalece el compromiso de las instituciones con la accesibilidad y la inclusión. Esta propuesta no ha sido todavía implementada, aunque si se han establecido las bases del diseño y planificación para su implementación en el marco de la edición actual del curso 2024-25 del proyecto Aprendizaje Servicio en el área de Biblioteconomía y Documentación: un proyecto generador de proyectos en bibliotecas, archivos y centros de documentación (Arquero et al., 2020; Arquero, 2024). Se estima que las limitaciones se podrían asociar a la evaluación del verdadero impacto que la puesta en práctica pueda tener y a los desafíos que puedan asociar en cuanto a la coordinación de las diferentes partes colaboradoras y su implicación en el proyecto. También se prevé que algunos de los materiales adaptados tengan que ser readaptados posteriormente para un mejor uso por parte de los usuarios, pues se aborda información diferente que requiere del testeo y validación por parte de personas con discapacidad auditiva que garanticen la adecuación de dichos materiales. Estas limitaciones pueden suponer una adaptación en el cronograma del proyecto planteado inicialmente.

No obstante, este proyecto sienta las bases de un trabajo potencialmente beneficioso para la comunidad. A largo plazo, se espera que este modelo de trabajo pueda ser replicado y ampliado, permitiendo que las bibliotecas continúen mejorando sus servicios y que los estudiantes sigan vinculando su formación académica con el servicio a la comunidad.

Agradecimientos

La participación de Elena Loreto Olmedo Pagés en este trabajo ha sido posible gracias a la financiación recibida como contratada predoctoral de la Universidad Complutense de Madrid en colaboración con Banco Santander y a su participación en el Proyecto de Aprendizaje-Servicio Complutense titulado ‘Aprendizaje Servicio en el área de Biblioteconomía y Documentación: un proyecto generador de proyectos en bibliotecas, archivos y centros de documentación’, cuya directora, Rosario Arquero Avilés, participa también en este trabajo. De igual modo, este estudio se ve respaldado por el Grupo IDEALab (https://www.ucm.es/idealab_grupodeinvestigacion), a quien se agradece el apoyo y la cooperación.

Sobre las autoras

Elena Loreto Olmedo Pagés es graduada en Historia por la Universidad de Alicante y con especialización en Bibliotecas por el Máster en Gestión de la Documentación, Bibliotecas y Archivos, ofertado por la Universidad Complutense de Madrid, universidad en la que continúa su formación en la investigación por medio del programa de Doctorado en Ciencias de la Documentación. Su tesis se enfoca en el diseño y desarrollo de productos y servicios bibliotecarios accesibles para personas con discapacidad auditiva. Puede ser contactada a través de elolmedo@ucm.es

Rosario Arquero Avilés es Doctora en Ciencias de la Información, licenciada en Documentación y diplomada en Biblioteconomía y Documentación por la Universidad Complutense de Madrid. Profesora Titular del Departamento de Biblioteconomía y Documentación de la Universidad Complutense de Madrid. Su perfil docente e investigador se relaciona con materias vinculadas a la dirección, planificación y gestión de proyectos y unidades de información y documentación, tanto en Grado, Máster como en Doctorado. Es directora del Grupo de Investigación IDEALab: Laboratorio de Innovación en Dirección, Evaluación y Administración de organizaciones y servicios de información y documentación, evaluado por la Agencia Española de Investigación del Gobierno de España y validado por la Universidad Complutense de Madrid. Es la responsable del Proyecto de Aprendizaje-Servicio Complutense titulado 'Aprendizaje Servicio en el área de Biblioteconomía y Documentación: un proyecto generador de proyectos en bibliotecas, archivos y centros de documentación' que, en el actual curso académico 2024-25, cumple su quinta edición con el respaldo y aprobación de la Oficina de Aprendizaje-Servicio (Unidad de Diversidad e Inclusión) de la Universidad Complutense de Madrid. Puede ser contactada a través de carquero@ucm.es

Referencias

Arévalo, J. y Quinde, M. (2023). La biblioteca como tercer lugar. Desiderata, (21), 106-112.

Arquero, R. (dir.) (2024). Aprendizaje-Servicio y Objetivos de Desarrollo Sostenible (ODS)-Agenda 2030: Laboratorio de ideas de proyectos en la formación de los futuros bibliotecarios, archiveros y documentalistas (Edición 2023-2024). Oficina de Aprendizaje-Servicio, Universidad Complutense de Madrid. https://hdl.handle.net/20.500.14352/108070

Arquero, R., Cobo, S., Marco, G. y Siso, B. (2020). Objetivos de Desarrollo Sostenible y Aprendizaje Servicio en la docencia universitaria: un estudio de caso en el área de Biblioteconomía y Documentación. Ibersid, 14(2), 13-24.
https://www.ibersid.eu/ojs/index.php/ibersid/article/view/4689

Batanero, C., De-Marcos, L., Fivera, L. F., Holvikivi, J., Hilera, J. R. y Tortosa, S. O. (2021). Improving Accessibility in Online Education: Comparative Analysis of Attitudes of Blind and Deaf Students Toward an Adapted Learning Platform. IEEE Access, 9, 99968-99982.
https://doi.org/https://doi.org/10.1109/ACCESS.2021.3095041

Calícrates (2018). Guía para hacer bibliotecas accesibles. Calícrates.
http://hdl.handle.net/11181/5442

Casals, I. (2006). La lectura fácil en las bibliotecas públicas. III Congreso Nacional de Bibliotecas Públicas: la Biblioteca Pública, espacio ciudadano (págs. 41-48).
https://travesia.mcu.es/server/api/core/bitstreams/d3f158ab-9fc8-43ff-b20d-264cb4fcafc8/content

Day, J. M. (2000). Lineamientos para los Servicios Bibliotecarios para Personas Sordas. IFLA.

Down España (2017). Guía Bibliotecas Accesibles. https://www.sindromedown.net/storage/2018/01/Gui%CC%81a-bibliotecas-accesibles.pdf

Folguerias, P., Luna, E. y Puig, G. (2013). Aprendizaje y servicio: estudio del grado de satisfacción de estudiantes universitarios. Revista de Educación, (362), 159-185.

González, F., Ochoa, A. y Guzón, J. (2022). Aprendizaje servicio en educación superior entre España y México. Alteridad, Revista de Educación, 17(1), 76-88. https://doi.org/https://doi.org/10.17163/alt.v17n1.2022.06

IFLA (2010). Guidelines for easy-to-read materials. IFLA. https://repository.ifla.org/handle/20.500.14598/637

IFLA y UNESCO (2001). Directrices IFLA/UNESCO para el desarrollo del servicio de biblioteca pública. IFLA, UNESCO. https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/publications/archive/the-public-library-service/pg01-s.pdf

Jiménez, L. (2020). Aps como instrumento de sensibilización hacia la discapacidad en educación superior. Revista INFAD De Psicología. International Journal of Developmental and Educational Psychology, 2(1), 305-312. https://doi.org/https://doi.org/10.17060/ijodaep.2020.n1.v2.1854

Litchke, L. G., Watt, T., Norton, C. L. y Finley, C. (2020). Impact of inclusive service-learning partnership on mental health among therapeutic recreation college students. Therapeutic Recreation Journal, 54(2), 189-206. https://doi.org/https://doi.org/10.18666/TRJ-2020-V54-I2-10034

Manzanares, G. (2016). Pautas para bibliotecas españolas al servicio de usuarios/as con diversidad funcional auditiva signantes. Boletín de la Asociación Andaluza de Bibliotecarios, (112), 124-141. https://aab.es/aab-boletin-112/

Martínez, J. A. (2007). L'accessibilitat a la informació en les biblioteques públiques: directrius per garantir la integració de tothom. BiD: textos universitaris de biblioteconomia i documentació, 18. https://bid.ub.edu/18marti1.htm

Mayor, D. y López, A. M. (2017). El aprendizaje-servicio como palanca para la inclusión educativa y social. Revista Educação Temas e Problemas, (17), 1-20.

Mendía, R. (2012). El Aprendizaje-Servicio como una estrategia inclusiva para superar las barreras al aprendizaje y a la participación. Revista Educación Inclusiva, 5(1), 71-82.

Mendía, R. (2016). El Aprendizaje-Servicio: una metodología para la innovación educativa. Convives, (16), 2-101.

Naciones Unidas (2024). Objetivos de Desarrollo Sostenible. Objetivos de Desarrollo Sostenible. https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

Plena Inclusión (2024). Diccionario Fácil. https://www.diccionariofacil.org/

Puig, J. (2019). Aprendizaje-servicio y educación en valores. Comunicación y Pedagogía: nuevas tecnologías y recursos didácticos, (319-320), 12-16.

Ríos, M. (2012). Inclusión del alumnado con discapacidad. Tándem: Didáctica de la educación física, (38), 7-18.

Sotelino, A., Santos, M. A. y Lorenzo, M. (2016). Aprender y servir en la universidad: Una vía cívica al desarrollo educativo. Teoría de la Educación, Revista Interuniversitaria, 28(2). https://doi.org/https://doi.org/10.14201/teoredu282225248

UNESCO (1994). Manifiesto de la UNESCO sobre la biblioteca pública. IFLA, UNESCO. https://www.ifla.org/wp-content/uploads/2019/05/assets/public-libraries/publications/PL-manifesto/pl-manifesto-es.pdf

United Nations (2015). Resolution adopted by the General Assembly on 25 September 2015. Transforming our World: The 2030 Agenda for Sustainable Development, A/RES/70/1 (21 October). https://sdgs.un.org/2030agenda

Uruñuela, P. M. (2018). La Metodología del Aprendizaje-Servicio. Aprender mejorando el mundo. Narcea.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Leveraging social circles and algorithmic processes in digital mental health tools for college students' stress management

Farhanuddin Fazaluddin Kazi and Jomara Sandbulte

DOI: https://doi.org/10.47989/ir30iConf47161

Abstract

Introduction. College students frequently face stress from academic pressures, social demands, and the challenges of daily life. Digital mental health (DMH) interventions have shown promise in helping manage these stressors. However, many students quickly abandon DMH tools, which limits their effectiveness. Previous research has identified social interaction as a key factor in improving user engagement with DMH tools. Still, most existing DMH tools have not been designed with these sharing practices in mind. Moreover, research on recommender systems (RSs) has mainly focused on single-user interactions, leaving a gap in the development of RSs for group settings, particularly in the health domain.

Method. To address this gap, we developed a technological probe that integrates social features with algorithmic processes to provide recommendations for both individuals and groups, aimed at supporting students' stress management efforts. We conducted a user study with 15 college students and their social circles, who interacted with the probe over the course of a day.

Results. Our findings revealed that students appreciated the opportunity to share daily activities with friends and felt motivated to engage in well-being-enhancing behaviours. Participants also gave valuable feedback on the stress management recommendations. We discuss and provide recommendations for future research.

Introduction

College students face stress from various sources, including academic demands, social pressures, and time management challenges (Choi, 2020, Rodgers et al., 2016). Digital mental health (DMH) tools hold significant potential to help students manage stress by incorporating personalization techniques (Cheung et al., 2019; Xu et al., 2021) through recommender system (RS) algorithms, which suggest relevant items to support users' health and well-being (Kelley et al., 2017; Rodgers et al., 2016). However, many users quickly abandon DMH tools, which limits their effectiveness to promote long-term health benefits (Wang et al., 2022). Thus, one of the primary design challenges for DMH tools remains increasing sustained user engagement (Lattie et al., 2020; Meyerhoff et al., 2022).

Research has shown that social interaction is a crucial factor in enhancing college students' engagement with DMH tools (Burgess et al., 2019; Meyerhoff et al., 2022; Park, 2018). College students are known to frequently share self-tracked data, such as exercise routines, mood logs, and study habits, with their peers (Gulliver et al., 2010, Lattie et al., 2020). Yet, DMH technology designs frequently overlook the social roles that support networks play in students' lives (Lattie et al., 2020; Park, 2018). Furthermore, most RS research has focused on single-user interactions, particularly in entertainment domains like movie recommendations (Delic et al., 2018, Tran et al., 2019). Research on health recommender systems (HRS) remains limited, especially in contexts where recommendations must direct multiple users working toward a shared goal (Alvarado et al., 2022; Coppens et al., 2023). There is a valuable opportunity to explore systems that not only address individual student's needs but also incorporate their social networks, aligning with their daily experiences and supporting their stress management efforts.

To address this research gap, we propose a DMH technology probe called '*FreeMind: Groups*', a social application designed for students to interact with peers and receive support in coping with stress. Our goal is to explore how a DMH design probe that incorporates college students' social circles and algorithmic processes could offer support for stress management, as well as to understand the benefits, challenges, and implications of such designs. To achieve this goal, we have examined the following research questions:

1. How can a design solution increase awareness of behaviours among students and their social circles?
2. How can health recommendation systems (HRS) be leveraged to promote collaboration in supporting students' self-care?
3. What concerns do people have regarding a) sharing experiences and b) algorithmic processes?

To answer our research questions, we conducted a user study with 15 college students and their social circle to understand individual's needs in a real-world setting and to inspire people to imagine the use of a new technology (Hulkko et al., 2004; Hutchinson et al., 2003). Participants interacted with our proposed system with a group of friends for a day. Our findings indicate that college students valued the opportunity to share their daily activities with friends and receive updates in return. Participants also reported feeling motivated to engage in positive activities for their personal well-being. Additionally, we present feedback from students on receiving recommendations for stress management and social support. Finally, we discuss the implications of these findings and offer recommendations for future research.

Related work

College students' stressor and social support

College students often experience significant transitions that can increase their stress levels (Nepal et al., 2022; Rodgers et al., 2016), potentially impacting their overall well-being (McLean et

al., 2022; Rodgers et al., 2016). While daily stressors are inevitable, certain factors can influence how stress affects students' well-being (Keyes, 1998). Social support plays a crucial role in helping individuals better manage daily stress by offering opportunities to adjust both individual and collective behaviours to promote mental health wellness (Lattie et al., 2020; Park, 2018). For instance, a friend can become a valuable partner for a student struggling with social pressure, providing unconditional support and guidance (Meyerhoff et al., 2022; Park, 2018). However, past conflicts and poor relationships can limit students' willingness to seek help and share information with their peers (Eisenberg et al., 2009; Park, 2018). Despite these challenges, leveraging social support can still offer significant benefits, such as reassurance and practical guidance in managing daily stressors (Feinberg et al., 2022; Park, 2018). Although previous studies have proposed strategies for incorporating social support into technology design (Burgess et al., 2019; Lattie et al., 2020), the social roles of individuals in college students' lives have often been overlooked (Lattie et al., 2020; Park, 2018). To bridge this research-to-practice gap, it is crucial to implement these strategies and evaluate their effectiveness under realistic conditions.

Inspired by prior works (Burgess et al., 2019; Burgess et al., 2020; Epstein et al., 2015; Hollis et al., 2017; Lattie et al., 2020; Morshed et al., 2019; Murnane et al., 2018), we developed a technology probe that integrates college students' social networks and algorithmic processes to support stress management. Our research makes two key contributions: a) providing insights into how the proposed system could aid college students in managing stress, and b) offering valuable guidance for designing interventions that encourage self-care practices.

Intelligent computing in health informatics

With advances in intelligent computing, systems enhanced by algorithm processes have shown great potential in supporting users' mental well-being (Hollis et al., 2017; Kazi & Sandbulte, 2023; Kim et al., 2022). For example, systems like Emotical (Hollis et al., 2017) predicts future moods and recommend healthy activities, while MindScope (Kim et al., 2022) predicts stress levels and provides explanations to aid user reflection. Building on prior works (Hollis et al., 2017; Kazi & Sandbulte, 2023; Kim et al., 2022), our study probe employed existing algorithms to provide both individual and group recommendations, while also incorporating explanations informed by prior literature (Afzal et al., 2018; Ehsan et al., 2021; Kim et al., 2022; Tran et al., 2019).

Group recommendations

Among various technologies, RSs are important in helping users efficiently navigate the overwhelming number of available options and information (Delic et al., 2018; Herzog & Wőrndl, 2019). However, these systems typically focus on serving individual users, often overlooking the potential to help groups. To address this gap, researchers have been increasingly implementing RSs in cooperative and social computing domains (Alvarado et al., 2022; Delic et al., 2018; Herzog & Wőrndl, 2019). In this context, group recommendation systems (GRS) offer valuable functionalities, such as selecting music that suits a group of people (Crossen et al., 2002). GRSs are most widely researched in entertainment, leisure, and event planning, where activities are often group-oriented and the system considers the preferences of the group (Alvarado et al., 2022; Herzog & Wőrndl, 2019; Pujahari & Padmanabhan, 2015). However, research in other areas, such as health, remains limited.

This study addresses this gap by using existing algorithmic processes to generate recommendations to support students' collaborative interactions around stress management. Those algorithms informed our design to a) better understand the user's inputs and b) enhance the quality of the recommendations based on the user's preferences (Alvarado et al., 2022; Pujahari & Padmanabhan, 2015).

Explanations in recommender systems

Prior studies have shown that explanations may influence user's behaviour, as they help users understand why certain items are recommended, thereby increasing the likelihood of user engagement (Sharma & Cosley, 2013; Tran et al., 2019). In GRS, explanations serve additional purposes compared to those in single-user systems. GRS must consider not only individual preferences but also the collective preferences of all group members to address fairness and resolve potential conflicts (Stettinger, 2014; Tran et al., 2019). An example of this is the Choicla group decision application, which supports various decision-making scenarios in a domain-independent manner (Stettinger, 2014). Building on these findings (Stettinger, 2014; Tran et al., 2019), we incorporated explanations within our recommendations. The explanations provided suggestions for coping with stress (G. Mitchell et al., 2021; Hollis et al., 2017; Kim et al., 2022) and utilized different explanation styles based on prior literature (Kouki et al., 2017; Liao et al., 2022; Sharma & Cosley, 2013).

Technology probe: design

We developed a technology probe called '*FreeMind: Groups*', a social application designed for students to interact with peers and receive support in coping with stress. *FreeMind: Groups* was built using Flutter (https://flutter.dev/) and the Django framework (https://www.djangoproject.com/).

Drawing on prior research (G. Mitchell et al., 2021, Kazi & Sandbulte, 2023, Kim et al., 2022, Lattie et al., 2020), the *FreeMind: Groups* system includes social features such as the ability to share photos, and the option to comment on and like posts, all designed to enhance users' social interactions (see Figure 1 - Left). Users receive automated recommendations in their feed from our RS named 'Zoha' (see Figure 1 - Right).

In addition, our probe includes a grouping function to reflect college students' social environments and roles (Burgess et al., 2019; Burgess et al., 2020; Kazi & Sandbulte, 2023; Park, 2018). Each group has a dedicated space to interact, addressing privacy concerns (Newman et al., 2011; Sandbulte et al., 2021). Users can view or join groups through a menu option that displays all their affiliated groups. Finally, users can select which group to post in, and each post within a group includes the group name and image to indicate its association (see Figure 1).

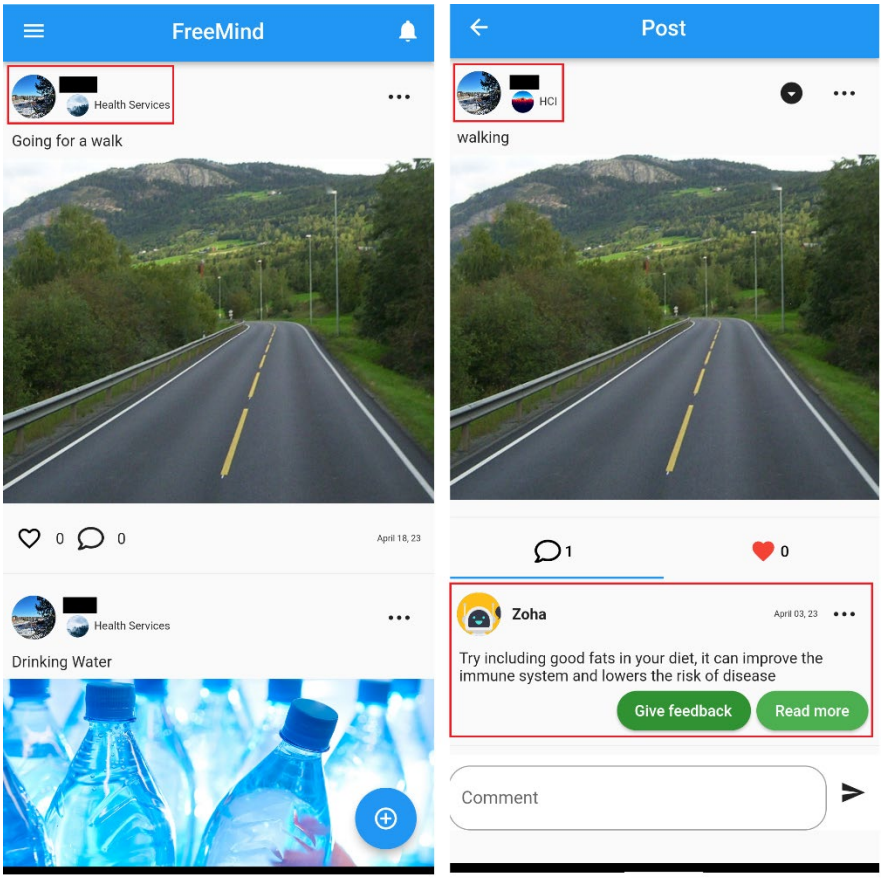


Figure 1. The *FreeMind: groups* probe. User feed with posts of daily activities (Left) and single-user recommendation sample based on survey data (Right).

Recommendation system

Using the recommendation model from (Kazi & Sandbulte, 2023), we generate single-user automated recommendations and deliver them as comments on the user's posts (see Figure 1, Right). The single-user recommendation is generated based on two data entries: a) the user's survey data and b) the user's caption from posts. Additionally, we have expanded the model in (Kazi & Sandbulte, 2023) by selecting the recommendation with the highest predicted rating, rather than randomly choosing from the top 10 highest-scoring options to enhance the quality of the suggestions.

Group-based recommendation system

We extend the work in (Kazi & Sandbulte, 2023) by developing group recommendations. Our system generates recommendations for group activities based on the preferences of all group members (Alvarado et al., 2022; Pujahari & Padmanabhan, 2015). These group recommendations are sent at specific times throughout the day, as determined by our research team.

To generate a group-based recommendation, the system uses collaborative filtering (Pujahari & Padmanabhan, 2015) to calculate the estimated rating of each activity with respect to each user in the group. For example, a group may have the following rating of an activity:

- User 1, Activity 1 = 5
- User 2, Activity 1 = 5
- User 1, Activity 2 = 3

Then, we combine the estimated ratings of activity given by each user (example: User 1, Activity 1 + User 2, Activity 1) resulting in a combined estimated rating of the activity for the whole group. Next, we select the activity with the higher rating. When the system recommends an activity, we included constraints to not allow to recommend the same activity repeatedly. The group recommendation is presented as a post without an image in the group feed to differentiate from the single-user recommendation (see Figure 2).

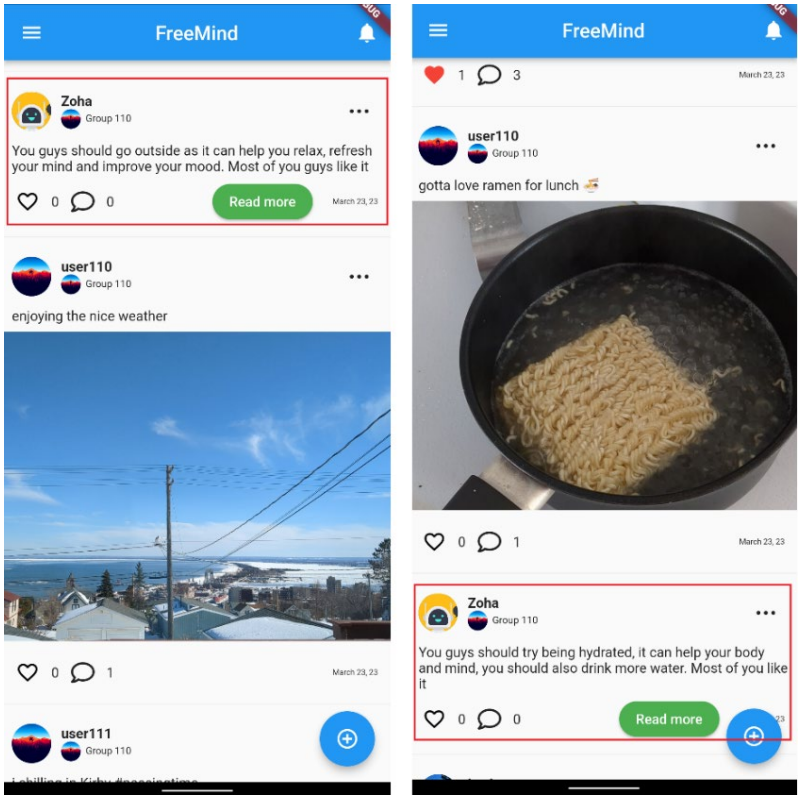


Figure 2. An example of a group recommendation suggesting physical activity (Left) and an example of a group recommendation suggesting a healthy behaviour (Right).

Recommendation explanations

We extend the recommendation explanation model presented in (Kazi & Sandbulte, 2023) by implementing three automated recommendations: a) knowledge-based (Liao et al., 2022), b) social-based (Sharma & Cosley, 2013), and c) hybrid-based (Kouki et al., 2017). The knowledge-based recommendation considers the post's caption and single-user survey preferences. Similarly, social and hybrid recommendations utilize the post's caption. Additionally, the social and hybrid recommendations incorporate the preferences of all group members to reflect the social context in the recommendation process.

We defined the following explanation templates for each recommendation based on prior literature to ensure they felt personalized and not '*robotic*.' Each recommendation includes an external link that directs students to our institution's website, where they can learn more about the benefits of the activity and explore available resources. Figure 3 illustrates how recommendation explanations are displayed within the application.

- **Knowledge-based**: this style presents a recommendation based on the benefits of the activity/strategy according to medical reports. For example: '*try going for a walk, since it strengthens muscles and bones, and improves physical energy and mood. Read more: [link]*'.

- **Social-based**: this style presents a recommendation based on social similarities within set of users (i.e., whether other group members like the activity too). For example: '*Try going for a walk, it is something your friends also like. Read more: [link]*'.
- **Hybrid-based**: this recommendation combines both knowledge and social styles. For example: '*Try going for a walk, since it improves physical energy and mood, and your friends like it. Read more: [link]*'.

Finally, an example of the group recommendation explanation template: '*You guys should go for a walk, since it improves physical energy and mood, and most of you like it. Read more: [link]*'.

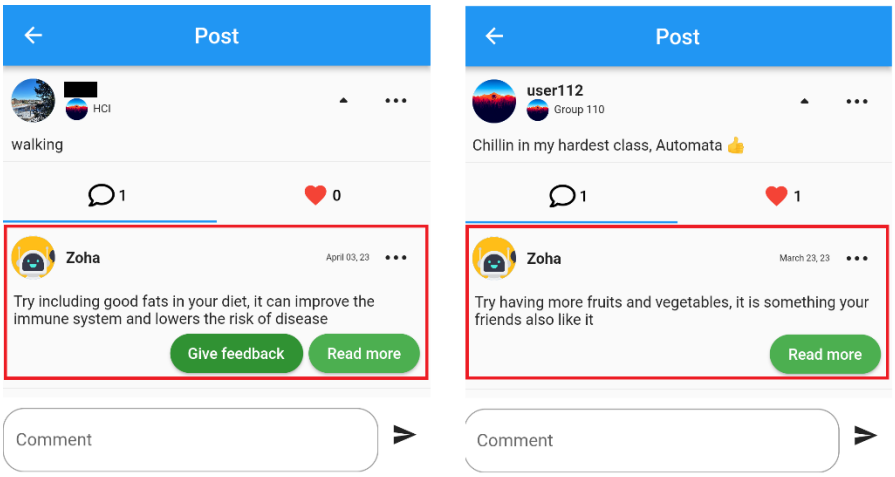


Figure 3. An example of a knowledge-based explanation style (Left) and an example of a social-based explanation style (Right).

Methods

This study aimed to explore how a DMH design probe that incorporates college students' social circles and algorithmic processes could offer support for stress management, as well as to understand the benefits, challenges, and implications of such designs. To achieve this, we conducted a user study involving 15 college students and their social circles in which they interacted with a technology probe for a day (approximately 8 hrs).

Participants recruitment

This research was approved by the Institutional Review Board (IRB) before any research activity. We recruited participants from various campus events and colleges, including the Computer Science (CS) graduate student's colloquium, distributed fliers around campus, and sent emails to student mailing lists.

Participants had to be 18+ years, currently enrolled in our institution, and were required to register in groups of at least three people. Interested individuals received an email detailing the study procedure, which involved using the app for a full day and providing feedback. Each participant was compensated with $30 for their contribution.

Study procedure

After receiving confirmation emails from all group members, we scheduled a study day for each group to participate. The study day was divided into two segments: a) an initial morning meeting and b) a final evening meeting. On the designated study day, we asked the entire group to meet in person with our research team during each session.

Initial morning meeting: We began by providing the participants with an overview of the application, and demonstrating basic tasks such as logging in and navigating the app. We also

explained the survey questions within the app and introduced the multi-group function. After the presentation, participants were asked to log into the app using pre-created usernames and passwords to simplify the process. We set up the accounts and added them to a group, enabling participants to communicate with each other within the app.

Once the application setup was complete, we discussed the participants' availability to determine a suitable time for the final meeting at the end of the day. Throughout the day, participants received at least four recommendations (i.e., one knowledge-based, one social-based, one hybrid-based, and one group based). After scheduling the final meeting, participants completed a pre-study survey including Perceived Stress Scale (Cohen et al., 1994), participants' social circle awareness level (Sandbulte et al., 2021), and demographics.

Final evening meeting: In the evening session of the study, we requested participants consent to record the interview session, and upon receiving verbal consent, we began the group interview. During the interview, we asked participants about their experiences using the application throughout the day. Some of the questions included: '*in your opinion, how can this app help you focus more on your self-care*?' and '*give an example of something new you learn about your friend during the study, if any*.'

Following the interview, participants completed a post-study survey including usability questions (Koester et al., 1998), health chatbot feedback, and participants' social circle awareness level (Sandbulte et al., 2021). At the end of the session, we provided each participant with compensation as an award for their contribution.

Data collection and data analysis

We employed a mixed-method approach for data analysis, integrating qualitative insights to explore quantitative findings more deeply (Wisdom & Creswell, 2013). The quantitative data was gathered using two methods: 1) Surveys were administered to collect information on individual perceived stress level and social circle awareness both before and after using the application; 2) Participant engagement was assessed through user activities like photos shared, comments, and likes, as well as social interactions within groups, including exchanged comments and likes.

For qualitative data, we conducted semi-structured interviews (Braun & Clarke, 2013). An inductive, qualitative analysis approach was used, where the research team open-coded the transcribed interview recordings to identify emergent themes. These themes were then discussed and refined to develop key, high-level insights. We focused on themes related to application use and interactions between students and their friends during the intervention.

Participant overview

A total of 15 students participated in the study (see Appendix A–Table 1). Among them, 8 participants identified as male and 7 as female. In terms of racial identity, 46% of the participants identified as White, 13% as Black or African American, and 40% as Asian. The age range was 20 - 32 years old (M = 22.8, SD = 3.46).

46% of the participants reported occasional feelings of being upset by unexpected events and 33% reported experiencing a lack of control over important aspects of their lives. Additionally, 40% of participants experienced nervousness and stress, highlighting the prevalence of stress-related issues (see Appendix A–Table 2).

On a more positive note, 53% of the participants felt confident in their ability to handle personal problems. However, challenges, such as feeling overwhelmed by difficulties, were also reported by 53% of the participants (see Appendix A–Table 2). These data underscore the complexity of well-being, revealing both resilience and vulnerability among participants.

Results

In this section, we summarize the key themes that emerged from our data analysis, which focused on participants' experiences with the application and aimed to address the study's research questions.

To systematically identify all participants, we developed a coding system. Each group was represented by the letter 'G' followed by a number, while individual participants within those groups were represented by the letter 'P' and a corresponding number. For example, group 1, participant 1 was coded as G1P1, and group 4, participant 2 as G4P2.

Friends as partners on stress management

Social ecosystems' awareness on daily stressors

In our study, we aimed to observe whether students and their social circle would increase awareness of each other's behaviors (RQ.1) and how recommendations could foster collaboration towards self-care (RQ.2). Our data revealed three primary ways in which the intervention influenced interactions within college students' social ecosystems.

Firstly, participants reported learning new things about each other while using the application. Even though the group members felt they already knew each other well, the intervention provided an opportunity to discover more about each other's daily routines. For instance, within group 4, participant G4P3 mentioned that they were all close friends who hung out together at least every two days. Yet, participant G4P2 was surprised to learn that G4P3 makes his bed every morning. Despite their familiarity, G4P2 appreciated the chance to learn more about each other's everyday habits: '*I learned that he [G4P3] made his bed which I didn't realize. I learned he had a good nap. It's interesting to see what you're doing*' (G4P2).

Another example of discovering new insights occurred in group 2. Participant G2P3 shared that she learned something unexpected about G2P1: '*I don't know G2P1 goes to gym. I meet her every day when all of a sudden, I saw a picture from G2P1 working hard at the gym.*' (G2P3) In response, G2P1 explained that going to the gym was a recent habit:

> *actually, I just started, I just decided to start. After Spring Break, I didn't have time because of homework and stuff. But then I'm like, 'today I have to go'. So, I just went and decided to post it* (G2P1).

During the interview, G2P3 noted that, although she knew her friends well, the intervention allowed them to gain new insights into each other's daily practices: '*I know them personally a lot, [but] I learned more specifically about their habits, likes, or dislikes, I guess. So, I learned a lot*' (G2P3).

Second, we noticed that participants reinforced their existing knowledge of their friends' behaviours. For instance, in group 2, G2P3 mentioned that she was already aware of G2P2's dislike for her Chemistry course. However, during the intervention, this knowledge was further solidified: '*most of her posts are related to the Chem course [then] I realized 'okay, she literally hates it* (G2P3).' In this case, the intervention didn't introduce new information but reinforced what was already known.

Finally, we observed that participants used this experience to clear up some misconceptions. For instance, within Group 1, participant G1P3 mentioned, '*I saw someone's jacket was purple, and I saw another one who really liked Valorant.*' (G1P3) Based on this post, G1P3 assumed that G1P2 liked the game Valorant. However, G1P2 clarified, '*No, I didn't like I am trying to like it* (G1P2).' This highlights how group interactions can sometimes lead to misunderstandings. In another case, G1P2 initially thought that G1P3 disliked Android studio but later discovered otherwise. G1P2 noted, '*I thought he actually dislikes Android, but it turned out he didn't hate it but and actually likes it. Yeah, I learned that*' (G1P2). To which G1P3 responded: '*lot of people think I hate Android, but I don't hate it*'(G1P3).

Leveraging social support to manage stress

During the study, we also noted some participants provided support by encouraging each other on healthy behaviors. For instance, in group 3' s case, G3P3 mentioned to see participant G3P1's post on doing physical activity and that motivated them to provide encouragement: '*I guess one thing I did was he was like working out in one of his posts and encouraged him, you know, like go get it or whatever, like encouraging him, I guess*' (G3P3). Reflecting on the study experience, participant G3P3 appreciated the opportunity to share daily activities with friends and receive updates from them: '*...it was nice to have everyone posting on. You see how everybody else is doing. Kind of keeps you engaged, you feel, I guess, better too*' (G3P3).

In addition, participants reported feeling motivated to engage in positive activities for their personal well-being to share their experiences. Many viewed this as a beneficial aspect of the study intervention. For example, in group 5, participant G5P2 noted that she decided to go to the gym and share it with her friends to receive support despite her busy schedule:

> *So today I was like, 'I need one post, so I have to go to the gym.' So, I just made myself to go for that. And I think if I just put myself in a situation that I need to do that because of sharing it with my friends and then they can give me feedback, then it would be a great way to make me do something (G5P2).*

Privacy considerations while sharing experiences

We also examined people's concerns about sharing personal experiences (RQ.3a). Participants highlighted the advantages of having a dedicated group space within the application instead of a public one. One key benefit was the reduction in stress and anxiety over what to post. For instance, participant G2P2 appreciated that the private space allowed her to share only with close friends, which reduced her stress: '*I like you're not posting it to everyone. It's just nice that it's more private like it's only two people that you're close to me' (G2P2).*

Participant G2P3 shared that having a dedicated space gave her the confidence to be vulnerable when discussing daily stressors. She also reflected on the anxiety she might feel if her posts were public:

> *if you're [thinking], what should I say? It's just I can be as vulnerable as I can with my own friends. But like for the friends whom I see once in a month, I really would be feeling threatened (G2P3).*

Participant G2P4 echoed this sentiment, highlighting the importance of group design and privacy: '*Yeah, I would say it's also privacy because sometimes you post something you don't want like others to see like just some specific people.*' (G2P4) These insights offer practical examples of how user needs can guide design decisions.

Participants also provided suggestions for improvements. For instance, participant G1P2 proposed adding an anonymous mode for sharing posts:

> *so, I think it is good to have friends in the same group, but it might be helpful to have anonymous option. So, people would post things like 'I failed on this exam.' Anyone could comment with usual text but [the post has] no name or picture (G1P2).*

He further elaborated on why an anonymous mode could encourage sharing: '*so that might help me feel better in some cases when sometimes is hard to share* (G1P2).'

Analysis of the recommendations' effectiveness to facilitate collaboration and promote well-being

In our study, we encouraged participants to reflect on their experiences with receiving recommendations for managing stress and facilitating social support (RQ.2). We also aimed to explore any concerns related to the algorithmic processes (RQ.3b).

Participants expressed overall satisfaction with the recommendations provided by the chatbot, finding them helpful for collectively improving lifestyles with friends (Appendix B-Table 5). For instance, participant G4P3 highlighted a benefit of receiving well-being reminders: '*...because it reminded me to do something that otherwise I wouldn't have* (G4P3).'

In group 3, all members reported receiving useful recommendations. Participant G3P1 shared: '*They [chatbot] told me to get water because I don't drink enough water during the day*' (G3P1) while G3P2 mentioned: '*It told me to eat my fruits and vegetables, which is good. I don't eat much of them. Get me on my toes.*' (G3P2). Additionally, G3P3 appreciated a specific recommendation, stating: '*I took a picture of me being in Stats class and It said, 'be open minded'. So, I thought that was pretty well.*' (G3P3)

However, participants felt that recommendations could be enhanced by making them more relevant to their posts rather than just their survey responses. For example, G5P2 observed: '*They're personalized now because it was relevant to what I chose in the surveys but not relevant to the post [image].*' (G5P2) Furthermore, participants suggested incorporating weather conditions based on location. For instance, G1P3 struggled with a recommendation due to poor weather: '*I mean, like It told you to go on a hike or go for a walk and it's like yesterday's weather, then you're probably gonna say I'm not going for a walk*' (G1P3).

Participants also showed a strong inclination to share health information with friends (Appendix B-Table 5). The system was reported to increase awareness of friends' lifestyles (Appendix B-Table 3) and facilitate health-related discussions among them (Appendix B-Table 3 and 5). For example, in group 5, participants mentioned they already had a WhatsApp group where they shared photos and chatted daily. However, the study intervention introduced a new topic of conversation — mental health. Previously, their discussions focused on parties and other activities, but the app prompted them to talk about health for the first time. Participant G5P1 shared:

> *now when I find something interesting, I send this for them and say, 'yeah, I found this really nice'. You can try that as well and they also give me some comments 'Oh yeah, let's try that together another time' we have the group to talk, manage anything, to ask question and talk about these specific things for our mental health, that is nice* (G5P1).

However, some participants felt that group recommendations could occasionally come across as rude, despite being based on observations. For example, G1P4 recounted a situation where a sarcastic comment he made led to an unintended personal recommendation from the bot:

> *I don't like how in one of the posts I made I said for the caption 'gotta enjoy Apple devices' and the bot said 'try being more kind' as it has been shown to boost mood and lengthen lifespan. I found that very personal to me* (G1P4).

Similarly, G4P2 commented on a group recommendation, noting that while they appreciated the suggestion to be optimistic, it might sometimes seem subjective or out of place:

> *I like the one about being optimistic, but it seems that one might be very subjective to put under someone's post because the bot might not really detect negativity, I guess. But as if it's just kind of a message for everyone in the group. I definitely like it as a recommendation (G4P2).*

These insights suggest that some recommendations may be better suited for general messages rather than comments on individual posts, as they could be perceived as targeting specific users, particularly when visible to the whole group.

Perceived benefits of recommendation explainability

Participants provided positive feedback on the recommendation system's ability to explain the reasoning behind its suggestions (see Appendix B – Table 4). For instance, participant G1P1 noted: '*I like the explanation because I just like to know why it is good to our health, although I don't want to do it right now. But it's good to know why it is good*' (G1P1).

Participants also preferred the hybrid approach used in the system, which combines both informational and social elements to enhance motivation for selfcare. Participant G4P1 said:

> *I like that there's kind of extra detail behind it that tells you how it can help you. And then I also do like knowing that my friends are doing this for wellness, well-being. So, it would make me want to do it too (G4P1).*

Participant G4P3 added that the '*read more*' option was beneficial for those who want additional information:

> *combination is definitely better I guess a lot for me, like you should do it because your friends are doing it and here's why and then you give like a short reason, and then you could extrapolate further if I decide to read (G4P3).*

However, some participants suggested improvements for the explanations, such as incorporating positive encouragement to support ongoing self-care. Participant G2P2 mentioned:

> *...even if it doesn't recommend based on what I was doing, but like if it just says, you're doing great, that comment is gonna make me smile more. I would say it doesn't particularly need to relate to something that I'm doing (G2P2).*

Similarly, participant G2P4 suggested that positivity could improve one's day, especially during stressful times:

> *so, I'm thinking maybe give positive notes like, oh, I have a great day. You're doing great ... I think you might like to make the person happy and like maybe the person is having a bad day, makes the person smile like* (G2P4).

Discussion

While our study involved a limited sample over a short period, the findings provide valuable insights into user experiences and perceptions of a social application designed to support stress management and well-being among college students. The following discussion highlights key themes that emerged from our data and explores their implications for design and future research.

Benefits of interventions leveraging social circles to support stress management

In this study, we identified three key findings: (1) increased awareness of everyday behaviours, (2) promotion of positive behaviours through social support, and (3) privacy considerations when sharing personal experiences.

First, the application significantly improved awareness within social groups about each other's daily stressors and habits. Increased awareness can encourage self-reflection, potentially leading

individuals to modify their own behaviours (Sandbulte et al., 2021). Our findings revealed that even among close friends, the intervention facilitated the sharing of new insights into daily routines and personal challenges that might not have been discussed otherwise (e.g., G4P2, G2P3). This aligns with previous research focused on creating spaces that strengthen social bonds (Epstein et al., 2015; Murnane et al., 2018; Park, 2018). Moreover, our study expands on prior work by showing that sharing experiences can reinforce existing knowledge and provide opportunities to clarify misconceptions, as demonstrated in the cases of G2P3 and G1P2. This suggests that increased awareness among participants not only fosters connection but also helps refine social perceptions, potentially reducing misunderstandings within social groups.

Second, the study emphasized the potential of social support in promoting positive behaviours. Participants reported feeling motivated to engage in healthy activities, like exercising, because they wanted to share these experiences with friends and receive encouragement (e.g., G5P2, G3P3). While previous research has shown that relationship conflicts can sometimes hinder the willingness to share (Eisenberg et al., 2009; Park, 2018), our findings demonstrate that positive social influence and support can significantly impact individual health behaviours. Social support can strengthen resilience to stress, and the quality of relationships plays a crucial role in shaping students' coping mechanisms and mental health outcomes (Keyes, 1998; Thieme et al., 2015). Group recommendations in this context, such as those highlighted by participant G3P4, further emphasized the importance of fostering a sense of community and shared responsibility for well-being. Future research could explore ways to maintain these partnerships for managing daily stress. Also, there is opportunity to explore different group dynamics such as non-collocated friends. Finally, future research in GRS could explore alternative media to help maintain group motivation since most recommendation systems rely on text messages (Cheung et al., 2019).

Lastly, privacy emerged as an important concern for participants. The strong preference for private group settings over public spaces highlights the importance of creating safe environments where users feel comfortable sharing personal experiences. Several participants, such as G2P2 and G2P3, reported reduced stress and anxiety when communicating in private settings which indicates the important role of privacy in fostering honest and open conversations. This aligns with existing research, which suggests that users are more inclined to disclose sensitive information in trusted, private spaces (Newman et al., 2011; Sandbulte et al., 2021). The suggestion to implement an anonymous mode (e.g., G1P2) further emphasizes the need for flexible privacy options to cater to varying levels of user's comfort. Addressing this need in future developments is essential, as customizable privacy settings can enhance user trust and promote a safer, more supportive experience.

Design considerations for developing effective recommendations

While most participants found the recommendations useful, our data revealed opportunities for improving personalization. Participants, such as G5P2, expressed a desire for recommendations that were more closely tied to their posts rather than only based on survey responses. This finding suggests an expectation for a more dynamic and context-aware system (Afzal et al., 2018; Cheung et al., 2019; Kazi & Sandbulte, 2023). Additionally, incorporating external factors such as weather conditions (e.g., G1P3) could further enhance the relevance and practicality of the recommendations. Previous research has explored algorithmic processes that consider users' contexts, like weather and physical activity preferences, to improve personalization (Afzal et al., 2018). Future studies could extend this contextual knowledge to better align the recommendations with real-time user's activities and environmental factors, which may increase user engagement and satisfaction.

Our study also addressed the gap in leveraging RS for group settings. Mixed responses to group recommendations where some participants found them too critical or too generalized (e.g., G1P4, G4P2) highlight the need to balance offering helpful advice while respecting individual's self-image

within a group. Similar challenges have been noted in other GRS domains, such as tourism (Delic et al., 2018; Herzog & Wőrndl, 2019). Future research could focus on exploring the optimal timing for delivering single-user versus group-based suggestions, to determine whether a recommendation is best suited for the whole group or not.

Finally, the positive reception of the recommendation system's explanations indicates that users value understanding the reasoning behind suggestions, even if they do not always act on them (e.g., G1P1). A hybrid approach combining informational and social elements was especially well-received, as it provided both context and motivation for the recommendations (e.g., G4P1, G4P3). However, feedback also suggests the explanations could incorporate positive reinforcement. For instance, simple motivational messages like *'you're doing great!'* could be integrated into the system's explanation to help sustain user's motivation.

Limitations

Our study introduces *FreeMind: Groups*, a social application designed to help students interact with peers and receive support in managing stress. We acknowledge the limitations of our brief user study and the preliminary nature of our findings, as the short duration may not fully capture the range of user behaviours. Future research should consider long-term interventions to address these limitations. Additionally, we acknowledge that our research is formative and informed by prior literature, and the limitation on relying on self-reported data. Moreover, similar to Newn et al., (2022), our study did not aim to evaluate different styles of explanation but rather the presence of explanations in general. We hope future studies can address these limitations.

Lastly, while we aimed to recruit a larger and diverse sample of students, we are aware that our sample size and participant group may not be representative of the general college student population. It is also possible that our sample may be affected by the Hawthorne effect, and social desirability bias. Still, our study's sample is consistent with the nature of this research (Braun & Clarke, 2013). Future studies should address these limitations, including a broader spectrum of stress levels and various types of friendship or group dynamics to gain deeper insights and propose innovative DMH interventions.

Conclusion

This study highlights the potential of a DMH tool that integrates social features and algorithmic processes to personalize the delivery of both individual and group recommendations aimed at supporting stress management for college students. Our findings indicate that college students valued the opportunity to share their daily activities with friends and receive updates in return. This interaction fostered a sense of connection and mutual support, which motivated participants to engage in positive activities for their personal well-being. Additionally, participants provided feedback on the recommendations they received for managing stress, highlighting the potential benefits of personalized, socially driven DMH tools. Finally, we offer recommendations for future research, emphasizing the need for further exploration into the design and implementation of socially integrated DMH tools that can effectively support college students' mental health.

Acknowledgements

The authors thank all participants for their valuable contributions to this study. This work was supported by the UMN Technology Commercialization under Early Innovation Fund Program.

About the authors

Farhanuddin Fazaluddin Kazi was a graduate student at the Department of Computer Science, University of the Minnesota -Duluth, Duluth, USA. He holds a master's degree in Computer Science. He can be contacted at kazi0021@d.umn.edu

Jomara Sandbulte is an Assistant Professor at the Department of Computer Science, University of the Minnesota -Duluth, Duluth, USA. She holds a master's degree in Computer Engineering and Ph.D. in Informatics from Pennsylvania State University. Her main research interests are in Human-computer interaction with AI systems, recommendation systems, and health informatics. She can be contacted at jsandbul@d.umn.edu

References

Abinash Pujahari and Vineet Padmanabhan. Group recommender systems: Combining user-user and item-item collaborative filtering techniques. In 2015 International Conference on Information Technology (ICIT), pages 148–152, 2015. https://doi.org/10.1109/ICIT.2015.36.

Amelia Gulliver, Kathleen M Griffiths, and Helen Christensen. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC psychiatry, 10(1):1–9, 2010.

Amit Sharma and Dan Cosley. Do social explanations work? studying and modelling the effects of social explanations in recommender systems. In Proceedings of the 22nd international conference on World Wide Web, pages 1133–1144, 2013.

Amra Delic, Julia Neidhardt, Thuy Ngoc Nguyen, and Francesco Ricci. An observational user study for group recommender systems in the tourism domain. Information Technology & Tourism, 19:87–116, 2018.

Andrew Crossen, Jay Budzik, and Kristian J Hammond. Flytrap: intelligent group music recommendation. In Proceedings of the 7th international conference on Intelligent user interfaces, pages 184–185, 2002.

Anja Thieme, Jayne Wallace, Thomas D Meyer, and Patrick Olivier. Designing for mental wellbeing: towards a more holistic approach in the treatment and prevention of mental illness. In Proceedings of the 2015 British HCI Conference, pages 1–10, 2015.

Christina Kelley, Bongshin Lee, and Lauren Wilcox. Self-Tracking for Mental Wellness: Understanding Expert Perspectives and Student Experiences, In Proceedings of the 2017 CHI conference on human factors in computing systems, page 629–641. Association for Computing Machinery, New York, NY, USA, 2017. ISBN 9781450346559. https://doi.org/2443/10.1145/3025453.3025750.

Corey Lee M Keyes. Social well-being. Social psychology quarterly, pages 121– 140, 1998.

Daniel A. Epstein, Bradley H. Jacobson, Elizabeth Bales, David W. McDonald, and Sean A. Munson. From 'nobody cares' to 'way to go!': A design framework for social sharing in personal informatics. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW' 15, page 1622–1636, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450329224. https://doi.org/10.1145/2675133.2675135.

Daniel Eisenberg, Marilyn F Downs, Ezra Golberstein, and Kara Zivin. Stigma and help seeking for mental health among college students. Medical Care Research and Review, 66(5):522–541, 2009.

Daniel Herzog and Wolfgang Wőrndl. User-centered evaluation of strategies for recommending sequences of points of interest to groups. In Proceedings of the 13th ACM Conference on Recommender Systems, pages 96–100, 2019.

Eleanor R. Burgess, Kathryn E. Ringland, Jennifer Nicholas, Ashley A. Knapp, Jordan Eschler, David C. Mohr, and Madhu C. Reddy. "I think people are powerful": The sociality of individuals managing

depression. Proc. ACM Hum.-Comput. Interact., 3(CSCW), nov 2019. https://doi.org/10.1145/3359143.

Eleanor R. Burgess, Renwen Zhang, Sindhu Kiranmai Ernala, Jessica L. Feuston, Munmun De Choudhury, Mary Czerwinski, Adrian Aguilera, Stephen M. Schueller, and Madhu C. Reddy. Technology ecosystems: Rethinking resources for mental health. Interactions, 28(1):66–71, dec 2020. ISSN 1072-5520. https://doi.org/10.1145/3434564.

Elizabeth L. Murnane, Tara G. Walker, Beck Tench, Stephen Voida, and Jaime Snyder. Personal informatics in interpersonal contexts: Towards the design of technology that supports the social ecologies of long-term mental health management. In Proceedings of the ACM on Human-computer Interaction 2 (CSCW), nov 2018. https://doi.org/10.1145/3274396

Elliot G. Mitchell, Elizabeth M. Heitkemper, Marissa Burgermaster, Matthew E. Levine, Yishen Miao, Maria L. Hwang, Pooja M. Desai, Andrea Cassells, Jonathan N. Tobin, Esteban G. Tabak, David J. Albers, Arlene M. Smaldone, and Lena Mamykina. From reflection to action: Combining machine learning with expert knowledge for nutrition goal recommendations. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI' 21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. https://doi.org/10.1145/3411764.3445555.

Emily G. Lattie, Rachel Kornfield, Kathryn E. Ringland, Renwen Zhang, Nathan Winquist, and Madhu Reddy. Designing Mental Health Technologies That Support the Social Ecosystem of College Students. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, page 1–15. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450367080. URL https://doi.org/10.1145/3313831.3376362.

Farhanuddin Fazaluddin Kazi and Jomara Sandbulte. Examining a social based system with personalized recommendations to promote mental health for college students. Smart Health, 28:100385, 2023. ISSN 2352- 6483. https://doi.org/10.1016/j.smhl.2023.100385.

Heidi Horstmann Koester, W McMillan, and MI Holland. Usability testing of software for assessing computer usage skills. In International Conference on Assistive Technology for People with Disabilities, 1998.

Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, St´ephane Coversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Bjőrn Eiderbăck. Technology probes: Inspiring design for and with families. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI' 03, page 17–24, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136307. https://doi.org/10.1145/642611.642616.

Ine Coppens, Luc Martens, and Toon De Pessemier. Analyzing accuracy versus diversity in a health recommender system for physical activities: a longitudinal user study. In Proceedings of the 17th ACM Conference on Recommender Systems, pages 1146–1151, 2023.

Jennifer Wisdom and John W Creswell. Mixed methods: integrating quantitative and qualitative data collection and analysis while studying patient-centred medical home models, 2013.

Jinkyung Choi. Impact of stress levels on eating behaviors among college students. Nutrients, 12(5):1241, 2020.

Jomara Sandbulte, Chun-Hua Tsai, and John M. Carroll. Working together in a Phamilyspace: Facilitating collaboration on healthy behaviours over distance. In Proceedings of the ACM on Human-computer Interaction, 5(CSCW 1), April 2021. https://doi.org/10.1145/3449198.

Jonah Meyerhoff, Rachel Kornfield, David C. Mohr, and Madhu Reddy. Meeting young adults' social support needs across the health behavior change journey: Implications for digital mental health tools. In Proceedings of the ACM on Human-computer Interaction., 6(CSCW2), pp. 1-33. https://doi.org/10.1145/3555203

Jordan Eschler, David C. Mohr, and Madhu C. Reddy. "I think people are powerful": The sociality of individuals managing depression. Proc. ACM Hum.- Comput. Interact., 3(CSCW), nov 2019. https://doi.org/10.1145/3359143.

Joshua Newn, Ryan M. Kelly, Simon D'Alfonso, and Reeva Lederman. Examining and promoting explainable recommendations for personal sensing technology acceptance. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 6(3), sep 2022. https://doi.org/10.1145/3550297

Kei Long Cheung, Dilara Durusu, Xincheng Sui, and Hein de Vries. How recommender systems could support and enhance computer-tailored digital health programs: a scoping review. Digital health, 5:2055207618824727, 2019.

Lavinia McLean, David Gaul, and Rebecca Penco. Perceived social support and stress: a study of 1st year students in Ireland. International Journal of Mental Health and Addiction, pages 1–21, 2022.

Mark E Feinberg, Jacqueline A Mogle, Jin-Kyung Lee, Samantha L Tornello, Michelle L Hostetler, Joseph A Cifelli, Sunhye Bai, and Emily Hotez. Impact of the covid-19 pandemic on parent, child, and family functioning. Family Process, 61(1):361–374, 2022.

Mark W. Newman, Debra Lauterbach, Sean A. Munson, Paul Resnick, and Margaret E. Morris. It's not that I don't have problems, I'm just not putting them on Facebook: Challenges and opportunities in using online social networks for health. In Proceedings of the ACM 2011 Conference on Computer Supported and Cooperative Work, CSCW '11, page 341–350, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450305563. https://doi.org/10.1145/1958824.1958876

Martin Stettinger. Choicla: towards domain-independent decision support for groups of users. In Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14, page 425–428, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681. https://doi.org/10.1145/2645710.2653365.

Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D. Abowd, and Thomas Plötz. Prediction of mood instability with passive sensing. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(3), pp. 1-21, 2019. https://doi.org/10.1145/3351233

Mengqi Liao, S Shyam Sundar, and Joseph B. Walther. User trust in recommendation systems: A comparison of content-based, collaborative and demographic filtering. In Proceedings of the 2022 CHI conference on human factors in computing systems, pages 1–14, 2022.

Muhammad Afzal, Syed Imran Ali, Rahman Ali, Maqbool Hussain, Taqdir Ali, Wajahat Ali Khan, Muhammad Bilal Amin, Byeong Ho Kang, and Sungyoung Lee. Personalization of wellness recommendations using contextual interpretation. Expert Systems with Applications, 96:506–521, 2018.

Oscar Alvarado, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. A systematic review of interaction design strategies for group recommendation systems. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2):1–51, 2022.

Pigi Kouki, James Schaffer, Jay Pujara, John O' Donovan, and Lise Getoor. User preferences for hybrid explanations. In Proceedings of the Eleventh ACM Conference on Recommender Systems, pages 84–88, 2017.

Sami Hulkko, Tuuli Mattelmäki, Katja Virtanen, and Turkka Keinonen. Mobile probes. In Proceedings of the Third Nordic Conference on Human-Computer Interaction, NordiCHI' 04, page 43–51, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138571. https://doi.org/10.1145/1028014.1028020.

Shannon Rodgers, Brittany Maloney, Bernd Ploderer, and Margot Brereton. Managing stress, sleep, and technologies: An exploratory study of Australian university students. In Proceedings of the 28th Australian Conference on Computer-Human Interaction, OzCHI' 16, page 526–530, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450346184. https://doi.org/10.1145/3010915.3010961.

Sheldon Cohen, Tom Kamarck, Robin Mermelstein, et al. Perceived stress scale. Measuring stress: A guide for health and social scientists, 10(2):1–2, 1994.

Subigya Nepal, Weichen Wang, Vlado Vojdanovski, Jeremy F Huckins, Alex da Silva, Meghan Meyer, and Andrew Campbell. Covid student study: A year in the life of college students during the covid-19 pandemic through the lens of mobile phone sensing. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. https://doi.org/10.1145/3491102.3502043

Sun Young Park. Social support mosaic: Understanding mental health management practice on college campus. In Proceedings of the 2018 Designing Interactive Systems Conference, DIS '18, page 121–133, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351980. https://doi.org/10.1145/3196709.3196787

Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, and Hwajung Hong. Prediction for retrospection: Integrating algorithmic stress prediction into personal informatics systems for college students' mental health. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI' 22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. https://doi.org/10.1145/3491102.3517701

Thi Ngoc Trang Tran, Műslűm Atas, Alexander Felfernig, Viet Man Le, Ralph Samer, and Martin Stettinger. Towards social choice-based explanations in group recommender systems. In Proceedings of the 27th ACM Conference on User Modelling, Adaptation and Personalization, pages 13–21, 2019.

Tong Wang, Wei Wang, Jun Liang, Mingfu Nuo, Qinglian Wen, Wei Wei, Hongbin Han, and Jianbo Lei. Identifying major impact factors affecting the continuance intention of mhealth: a systematic review and multi-subgroup meta-analysis. NPJ Digital Medicine, 5(1):145, 2022.

Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: Towards social transparency in ai systems. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI' 21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. https://doi.org/10.1145/3411764.3445188.

Victoria Hollis, Artie Konrad, Aaron Springer, Matthew Antoun, Christopher Antoun, Rob Martin, and Steve Whittaker. What does all this data mean for my future mood? actionable analytics and

targeted reflection for emotional well-being. Human–Computer Interaction, 32(5- 6):208–267, 2017. https://doi.org/10.1080/07370024.2016.1277724.

Virginia Braun and Victoria Clarke. Successful qualitative research: A practical guide for beginners. Sage Publications, UK: London, 2013. Eleanor R. Burgess, Kathryn E. Ringland, Jennifer Nicholas, Ashley A. Knapp,

Xuhai Xu, Prerna Chikersal, Janine M. Dutcher, Yasaman S. Sefidgar, Woosuk Seo, Michael J. Tumminia, Daniella K. Villalba, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Afsaneh Doryab, Paula S. Nurius, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. Leveraging collaborative-filtering for personalized behaviour modelling: A case study of depression detection among college students. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(1), mar 2021. https://doi.org/10.1145/3448107.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Appendix

Appendix A: pre-study survey results

Group ID	Participant ID	Age	Gender
G1	P1	20	Male
	P2	32	Male
	P3	28	Male
G2	P1	28	Female
	P2	24	Female
	P3	22	Female
G3	P1	22	Male
	P2	21	Male
	P3	21	Male
G4	P1	21	Female
	P2	21	Male
	P3	21	Male
G5	P1	21	Female
	P2	21	Female
	P3	20	Female

Table 1. Participants' groups, age, and gender

	Question	Never	Sometimes	About half the time	Most of the time	Always
1	In the last month, how often have you been upset because of something that happened unexpectedly?	13.34%	46.67%	13.34%	20%	6.67%
2	In the last month, how often have you felt that you were unable to control the important things in your life?	20%	33.34%	13.34%	20%	13.34%
3	In the last month, how often have you felt nervous and stressed?	13.34%	40%	6.67%	20%	20%
4	In the last month, how often have you felt confident about your ability to handle your personal problems?	6.67%	13.34%	13.34%	53.34 %	13.34%
5	In the last month, how often have you felt that things were going your way?	6.67%	40%	13.34%	40%	0.00%
6	In the last month, how often have you found that you could not cope with all the things that you had to do?	20%	53.34%	20%	6.67%	0.00%
7	In the last month, how often have you been able to control irritations in your life?	13.34%	13.34%	33.34%	33.34 %	6.67%
8	In the last month, how often have you felt that you were on top of things?	6.67%	40%	13.34%	33.34 %	6.67%
9	In the last month, how often have you been angered because of things that happened that were outside of your control?	6.67%	46.67%	13.34%	26.67%	6.67%
10	In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?	26.67%	53.34%	0.00%	20%	0.00%

Table 2. Pre-study survey perceived stress scale (PCS) questions. Likert scale 0 to 5 (0= Never, 5=Always), n=15

Appendix B: post-study survey results

	Questions	Min.	Max	Mean	SD	Variance
1	I am more aware of my friend's healthy behaviors after participating in this study.	3.00	5.00	4.27	0.77	0.60
2	I feel happier about my friend's healthy lifestyle after participating in this study.	3.00	5.00	4.13	0.81	0.65
3	I feel closer to my friend after participating in this study.	1.00	5.00	4.13	1.09	1.18
4	I give more feedback to my friend about his/her healthy lifestyle after participating in this study.	0.00	5.00	3.67	1.30	1.69
5	I receive more emotional support from my friend after participating in this study. For example, I receive messages of encouragement to go exercise.	2.00	5.00	4.07	1.00	1.00
6	I receive more tangible support from my friend after participating in this study. For example, He/She goes exercising with me.	2.00	5.00	3.87	1.15	1.32
7	I receive more informational support from my friend after participating in this study. For example, He/She share articles about healthy living with me.	0.00	5.00	3.87	1.31	1.72
8	I feel I have a very healthy lifestyle.	1.00	5.00	3.87	1.20	1.45

Table 3. Survey awareness questions. Likert scale 0 to 5 (0= Strongly disagree, 5=Strongly agree), n=15

	Questions	Min.	Max.	Mean	SD	Variance
1	It was very easy to learn how to use the FreeMind: groups application	4.00	5.00	4.80	0.40	0.16
2	Using the FreeMind: groups application was a very pleasant experience.	4.00	5.00	4.60	0.49	0.24
3	I would like to use the FreeMind: groups application outside the study	3.00	5.00	4.20	0.75	0.56
4	I would really dislike to use the FreeMind: groups application daily	0.00	5.00	0.93	1.28	1.64
5	The recommendations were confusing	0.00	5.00	1.60	1.58	2.51
6	The recommendations were helpful	1.00	5.00	3.13	1.31	1.72
7	The recommendations were relevant	0.00	5.00	2.80	1.47	2.16
8	I read more about the recommendation activity by going to the link provided	0.00	5.00	3.07	1.53	2.33
9	I liked the way that the recommendation was presented	2.00	5.00	3.60	0.95	0.91
10	I disliked the way that the recommendation was presented	0.00	5.00	1.67	1.49	2.22
11	I was able to effectively interact with the FreeMind: Groups application	4.00	5.00	4.73	0.44	0.20

Table 4. Survey usability questions. Likert scale 0 to 5 (0= Strongly disagree, 5=Strongly agree), n=15

	Questions	Min.	Max.	Mean	SD	Variance
1	The health chatbot was providing good health information to me.	4.00	8.00	6.87	1.02	1.05
2	The health chatbot was useful for improving my lifestyle.	4.00	8.00	6.67	0.87	0.76
3	I like the health information recommended by the health chatbot.	4.00	8.00	6.80	1.11	1.23
4	I am convinced by health information recommended by the health chatbot.	4.00	8.00	7.07	1.06	1.13
5	The health chatbot explained why the health information was recommended to me.	4.00	8.00	6.73	1.06	1.13
6	I would share the health information recommended by the health chatbot with my friends.	4.00	8.00	7.07	1.06	1.13
7	The health chatbot helped me to improve my lifestyle with my friends.	4.00	8.00	6.80	1.11	1.23
8	The health chatbot made me more aware of my friend's lifestyle.	4.00	8.00	7.20	1.05	1.09
9	The health chatbot helped me to talk about health with my friends.	4.00	8.00	6.67	1.25	1.56

Table 5. Survey questions on health chat bot feedback. Likert scale 0 to 5 (0= Strongly disagree, 5=Strongly agree), n=15

Information Research - Vol. 30 No. iConf (2025)

The role of paralinguistic digital affordances (PDAs) in online mental health communities (OMHCs) in Saudi Arabia

Sara Alrashed, Peter A Bath, and Kushwanth Koya

DOI: https://doi.org/10.47989/ir30iConf47158

Abstract

Introduction. Depression is a major contributor to disability and will be a leading cause of global disease burden by 2030. There is a high depression rate in Saudi Arabia. Online Mental health communities (OMHCs) are important for people with mental health conditions who are concerned about the stigma associated with these conditions. Paralinguistic digital affordances (PDAs) are a useful social affordance enabled by social media. This study aimed to understand the role of using PDAs in Saudi OMHCs.

Method. Qualitative methods were employed to analyse a sample of 1331 posts from two Saudi OMHCs.

Analysis. Data were thematically analysed using Braun and Clarke's method.

Results. PDAs fostered receiving emotional and informational support and were used to communicate frustrating emotions. A variety of emojis was used to provide encouragement to other users in response to their posts and conveyed different forms of support. PDAs were also used to show empathy and to acknowledge other people's feelings and to reassure them. However, on occasion, PDAs were used in a negative way.

Conclusion. PDAs provide additional emotional and informational support to encourage people with depression in OMHCs. Future research could interview OMHC users to develop a deeper understanding of how PDAs provide support.

Introduction

Mental health is considered to be one of the fundamental aspects of overall health and well-being, and is affected by a range of mental illnesses, including depression, anxiety, and stress. Depression is one of the most prevalent, complicated, and harmful mental health illnesses (Durbano et al., 2024; Mclaughlin, 2011). Globally, it is a significant cause of disability and is expected to be the leading cause of disease burden by 2030 (Lopez & Murray, 1998; Rostek & Gamble, 2020). The World Health Organization (WHO, 2023) estimates that approximately 3.8% of the world's population is currently experiencing depression, and the number of people affected is rising. It is considered a global public health concern that affects people of all ages and is an underlying reason for self-harm and suicide (Nusawat & Leelasantitham, 2024; Xiao et al., 2022).

Saudi Arabia is similar to many other countries in that depression is considered one of the leading mental health illnesses. Alhabeeb et al. (2023) estimated that the prevalence of major depression disorder (severe depression) in Saudi Arabia is 12.7%. The prevalence of depression varies considerably across the 13 regions of Saudi Arabia from 8.2% (Qassim and Hail) to 19.5% (Najran). Recently, Odah et al. (2024) conducted a study in southwestern Saudi Arabia and found the average prevalence rate of depression to be 68.1%. The distribution of this prevalence was mild (28.2%), moderate (21.9%), moderate to severe (12%), and severe (6%). The high depression rates are a major public health concern in Saudi Arabia.

Social media has become an important part of people's lives (Nusawat & Leelasantitham, 2024). Social media significantly improves the capacity of user-generated content and allows people to communicate with others at any time. It provides users with a virtual space to communicate with others in a relatively fast and inexpensive way. Social media has become a convenient means for sharing and exchanging health-related information with others (Lin & Kishore, 2021; Yoon & Syn, 2023). Individuals can share health topics, advice, as well as their personal health experiences (Hargreaves et al., 2018; Lovatt et al., 2017). Social media facilitates collaboration and communication among patients with specific health conditions.

Social media has transformed how individuals communicate, interact, and how they perceive social support. With various features and functionalities, it provides opportunities (affordances) for individuals to access and potentially engage in social support. Although several studies have examined health information behaviour within the context of social media (Afful-Dadzie et al., 2023; Kanchan & Gaidhane, 2023), little attention has been paid to the effect of its features in the health context (Wu et al., 2024; Zhang et al., 2018). Exploration of the integration of social media features and user actions, i.e., social media affordances related to health information, could offer a potential understanding of the complex role of social media in the health context (Wu et al., 2024). Additionally, Nattembo (2022) indicated that few studies have investigated how social media affordances influence mental health. Addressing this gap is critical for developing a deeper understanding of the complexities of social media use in the context of mental healthcare.

One of the affordances of social media is paralinguistic digital affordances (PDAs), which refer to symbols in social media messages that facilitate interaction between individuals without determining a specific language; examples of these symbols include 'emojis' (e.g., 😊), '*liking*' a message (e.g., 👍) (Hayes et al., 2016b), stickers or avatars. Additionally, PDAs can include the number of emojis and likes attributed to a post. PDAs are not only about emojis or emoticons and how users interact with them but also include the context in which users apply them.

Few studies have investigated the effect of emojis in text, as well as the various text and emoji combinations, on users' mental state (Pfeifer et al., 2022). The underlying mechanisms of the role of PDAs could be illuminated by social support theory, since social support has been indicated as a main driver of social media activities, including for the provision of mental health information.

In Saudi Arabia, seeking and accessing online health information has become increasingly common. According to Madhi et al. (2020), online health information is often accessed via social media platforms such as X (formerly Twitter), Snapchat, and online communities. Online health communities in Saudi Arabia have been only recently initiated, yet people have already started to use them to share their experiences. However, to the best of the authors' knowledge, no study has investigated the role of PDAs on social media in the Saudi health sector. Therefore, the aim of this paper is to investigate the role of PDAs in perceiving social support in online mental health communities (OMHCs) in Saudi Arabia.

The following section first provides a theoretical background of the affordance lens and social support theory. Then, through a qualitative study design using online posts, the role of PDAs in receiving social support in OMHCs is investigated. Finally, the paper discusses the findings and implications.

Literature review

Social media affordances

The term 'affordance' was first coined by Gibson (1977) in the context of ecological psychology and analysed the relationship between the environment and animals. Generally, affordance refers to opportunities for action that a person (subject) perceives while interacting and communicating with an object or artefact. Therefore, affordances are neither subjective (actors) nor objective (artefacts); they cross the dichotomy between subjective and objective. However, Norman's (1988) definition refers to the properties that include both perceived and actual things that identify how things might be used. Ronzhyn et al. (2022) debated the differences between Gibson's (1977) understanding of affordances, which separates the actor from the action, and Norman's (1988) understanding of affordances, which stand or exist if they are perceived by a subject (actor). The main theme of all the extant definitions is that affordances involve an object that facilitates an action (Leidner et al., 2018).

As the term has developed, researchers have proposed technology affordance theories. Technology, an object, can provide numerous affordances. Technological affordances can be defined as opportunities for action offered to an individual by technological artifacts (Argyris & Monu, 2015; Gibson, 2014). There is a relationship between individuals and the material properties of a technology (Vaast et al., 2017). Caliandro and Anselmi (2021) proposed that technological affordances reside in a group of contextual limitations and props that characterize technology use. Since they are relational, perceptional, and contextual, the same technologies, even the same technological properties, can lead to different affordances (Ronzhyn et al., 2022; Treem & Leonardi, 2013). Moreover, how these affordances apply to different users will vary (Steinert & Dennis, 2022).

Social media affordances are considered a stream from technological affordances. Ronzhyn et al. (2022) defined social media affordance as how an individual can perceive actual or imagined properties from social media that are the result of the technological, social, and contextual relationship. This relationship enables and shapes the specific usage of platforms. The design of social media features allows individuals to share knowledge that increases user engagement and potential perceiving social support (Lin & Kishore, 2021; Majchrzak et al., 2013; Treem & Leonardi, 2013).

Various studies have incorporated affordance as a valuable theoretical lens for investigating the effects of media platforms on society (e.g., Vaast et al., 2017). Four types of social media affordances were proposed by Treem and Leonardi (2013) in an organizational context. The first is visibility, which means that employees can make information and expertise more visible to others. The second is editability, which means messages can be created, edited, and improved before being published. The third is persistence, which enables employees to understand their past and present workplaces and their co-workers. The fourth is association, which includes how social media

enables employees to exchange and sustain relationships between them, as well as between employees and content experience-related information. These affordances contribute to noticeable knowledge exchanges. Lin and Kishore (2021) proposed three further social media affordances in the health context: community co-creation, social learning, and social relationships. Community co-creation in social media includes when patients seek or offer information, such as asking questions, and this can foster information support. The affordance for social learning is the process by which individuals with common interests or concerns can communicate, interact, observe, and learn from each other (Conte & Paolucci, 2001). For example, patients can observe others via online groups to learn how to cope with a specific illness, which can foster experience support. The third affordance is for social relationships, which means developing and maintaining social relations between users in social media, which can foster emotional support. These affordances contribute to social support. Lu et al. (2024) proposed three digital affordances that can be applied to social media or other technology and were applied in a health crisis context. These affordances were sharing information, sharing knowledge and socializing, and they can enhance the ability to receive social support.

Evidence from prior studies has revealed that social media affordances enhance user engagement through their properties or features, such as sharing information tasks. In turn, individuals become capable of creating value and expanding resources in their preferred ways. Moreover, Karahanna et al. (2018) found that social media features provide or facilitate social media affordances that generate outcomes, which can be knowledge sharing or perceived social support. A further important social media affordance is PDAs (Sun et al., 2021), which is discussed in the following section.

Paralinguistic digital affordances (PDAs)

PDAs are phatic communications that offer diverse social media functions or features, such as 'likes' and 'emojis'. One of the main motives for using PDAs in social media is to offer social support (Carr et al., 2016; Hayes et al., 2016b; Lyubenko, 2023; Sun et al., 2021). Several studies have investigated PDAs as a form of perceived social support. For example, Rawlings (2023), using a quantitative approach, found that receiving likes on Facebook is associated with social support. Johnson et al. (2020) collected posts related to infertility treatment and used content analysis to examine emoji use and social support on Instagram. One of their findings was that users received emotional support when they engaged in self-disclosure.

Various studies have highlighted how affordances are a relational property because different individuals may possess varying abilities, leading them to perceive and use social media platforms in different ways (Gibson, 2014; Page et al., 2022). Moreover, individuals seek various aspects of support on social media, depending on platform features and the user services (Hayes et al., 2016a). Even though affordances are similar in online health communities and general-purpose social media, health-focused online communities must consider the underlying contextual differences (Smith et al., 2023). Compared to other online health communities, posts in OMHCs have been shown to have more expressions of emotion (Gu et al., 2023). Thus, it is important to investigate PDAs in the context of OMHCs to gain a better understanding of how and why they are used in these platforms. This study examined how people with depression incorporate and use PDAs in their responses to OMHCs and investigated the role of PDAs in the relatively new context of OMHCs in Saudi Arabia.

Several studies have focused on investigating PDAs from a microscopic point of view, such as the frequency of PDA use (Marengo et al., 2021; Tang, 2022). The collective aspect of PDA use has been largely overlooked (Sun et al., 2021). Therefore, OMHCs feature the combination of emojis with text for social support extraction; hence, an analysis of both written text and visual symbols is needed to obtain better insights into people's emotions and how people engage in such communities. It is also necessary to investigate the outcomes from using PDAs and social support for people with

depression, as these outcomes do not always align with what they were likely intended to be. Thus, the study aimed to understand the role of PDAs in potentially benefiting or harming people with depression in OMHCs.

Social support theory (SST)

SST was initially developed in the mid-1970s and early 1980s. Formulated by Cobb (1976), SST describes information that can lead an individual to believe that they are loved and respected and that they belong to a community that provides mutual obligations. At its core, SST posits how social support might help others effectively, with individuals offered assistance available on a network. SST provides a person's perceptions of a situation and involves respect, concern, or responsible regard for others. Cobb (1976) posited that social interaction can help individuals develop mental health resilience to negative emotions. Thoits (1982) indicated that the concept of social support is the degree to which individual social needs are fulfilled via interaction with others. Here, social support emphasizes the exchange of resources over interpersonal connections (Cohen & Hoberman, 1983). Subsequently, Shumaker and Brownell (1984) highlighted that the concept involves an exchange between at least two users to improve the well-being of the receiver (of the support). According to SST, the support that users receive (e.g., receiving care and help from the community) may improve their sense of competence (Hasan et al., 2023). The main role of social support concerns people's ways of encouraging and helping others to experience more confidence in social settings. Therefore, establishing trust among members and addressing members' psychological needs are important (Islam et al., 2018).

Generally, the existing literature has classified social support into various taxonomies. Information and emotional support are the main types of social support. Emotional support includes listening to others, offering encouragement, and showing trust, empathy, understanding, and concern (Johnson & Lowe, 2015; Krause, 1986; Zeitouni & Milstein, 2017), for example, making patients feel loved, valued, and empathized with. Informational support includes providing assistance in the form of knowledge, transmitting information, guidance, and advice, and helping others to make decisions and to minimize uncertainty, for example, when patients discuss health topics such as symptoms or how to cope with disease. This paper investigates how people with depression use PDAs to foster emotional and informational support.

Method

This qualitative study explored PDA use for people with depression in OMHCs in Saudi Arabia. Data were collected using online posts from two OMHCs in Saudi Arabia. There are various reasons for utilizing online posts, for example, to explore, elaborate, and enhance the understanding of the role of PDAs (such as how patients use these PDAs in seeking or offering social support in OMHCs). Collecting and analysing online posts also allows investigating patients' experiences and views without any explicit involvement or implicit interference: this approach has been used successfully in previous studies (Hargreaves et al., 2018; Lovatt et al., 2017; Xu & Li, 2023).

Study setting

Amer is one of the first OMHCs in Saudi Arabia, and it is approved by the Ministry of Health. The platform has 13 forum boards, such as '*the past*', '*self-development*', and '*finding balance*', and various features that enable people with depression to write posts, adding text, emojis, and hashtags. People can read and reply to a thread or another comment in the same thread by either writing a comment and adding emojis or adding any PDAs, including likes, emojis, or other visual icons. The second OMHC is the Telegram public group, run by a moderator (not a therapist). This community discusses depression. Figure 1 presents examples of emojis used on Amer and the Telegram public group. The posts are public and written in Arabic. Additionally, people can anonymize themselves. The authors contacted the gatekeepers for their approval and obtained ethical approval, as described in the next section.

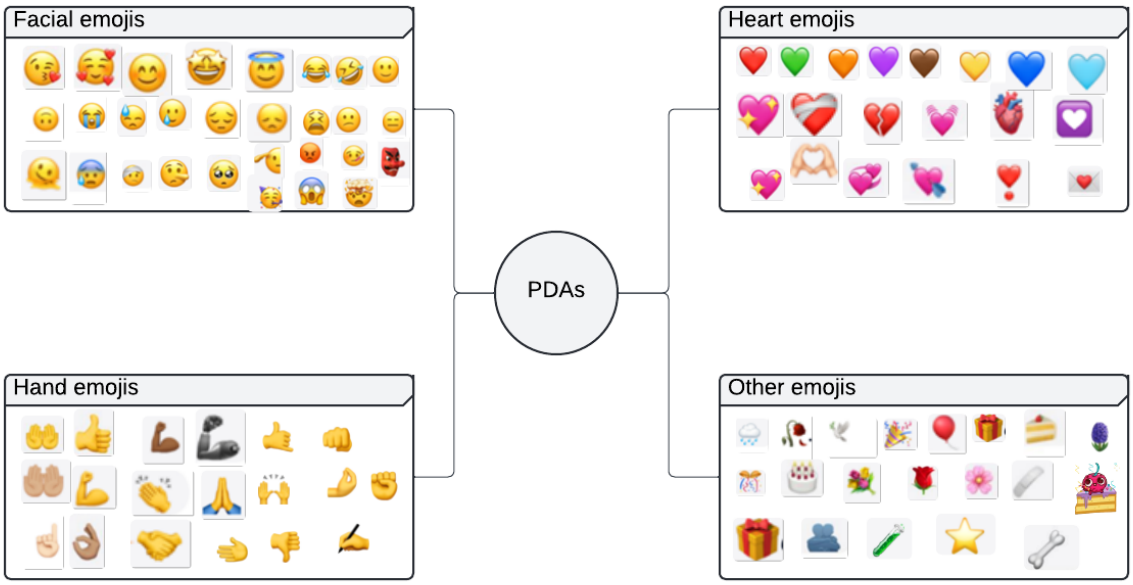


Figure 1. Examples of paralinguistic digital affordances (PDAs) used in Saudi OMHCs

Ethics approval

Ethics approval was obtained from the University of Sheffield in May 2024 (application no.: 059127). We received permission from OMHC gatekeepers to access and analyse the data, which are posts available to the public. The study followed the University of Sheffield research ethics policy (University of Sheffield, 2023).

Confidentiality

One of the OMHCs has been renamed Amer to maintain its anonymity. Additionally, all details of posts were anonymized to ensure the confidentiality and privacy of the posts. Posters' usernames were removed and replaced with pseudonyms. The quotations used in this study were translated into English. For posts with identifying markers after translation, the researcher changed the words of the posts but kept the same meaning; however, in practice, this was rarely necessary.

Sampling strategy

For Amer OMHC, the forum boards were selected purposively based on their richness and relevance to the study's topic. Four boards were selected: self-development, general, recovery stories, and health. A sample of the data (1331 posts) related to threads in those boards and the Telegram public group was collected. The posts were sampled to reflect patients' beliefs and provide different views. Data were stored and managed using ATLAS.ti. Inductive and deductive thematic analyses were conducted.

Data preparation and analysis

After collecting the posts from the two OMHCs, the researcher (SAR) prepared the data for analysis. They downloaded the threads and stored them in WORD files prior to being anonymized. The files were converted to PDFs and uploaded into ATLAS.ti. The posts were analysed in Arabic. Materials that needed to be quoted or discussed in detail with the other members of the research team were translated into English.

Thematic analysis was conducted to analyse the collected posts (Braun & Clarke, 2006). Thematic analysis is a frequently used method of analysis in social and health care studies (Pope & Mays, 2020; Swain, 2018), and several studies have used it in the online health community context

(Hargreaves et al., 2018; Lovatt et al., 2017). Thematic analysis is not concerned with a particular philosophy or epistemology and, therefore, can be used in various epistemological methods (Braun & Clarke, 2006). In addition, it has the flexibility to identify relations (patterns) within data related to participants' views, experiences, and behaviours (Clarke & Braun, 2017).

The collected posts were analysed, following Braun and Clarke's (2006) guidelines, as shown in Figure 2. The posts were read several times, and any ideas formulated while reading were written down. Data features that addressed the research topic were then coded. The codes were shared within the research team for discussion and agreement. The codes were then organized thematically into meaningful groups; codes that appeared consistently and were similar were considered inductive themes.

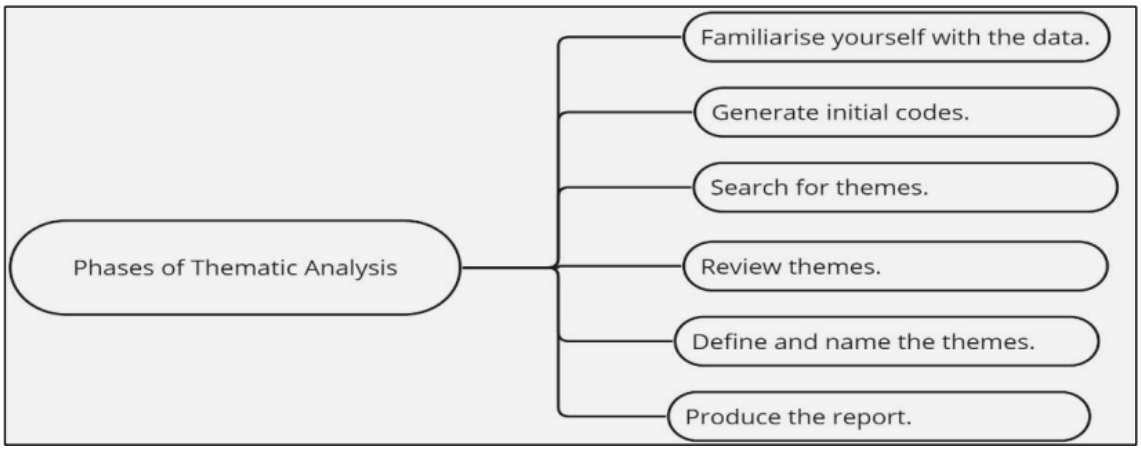


Figure 2. Phases of thematic analysis (adapted from Braun and Clarke, 2006)

Results

After analysing the data, three main themes emerged relating to the use of PDAs: emotional support, informational support, and PDA as a harmful tool. The integration of these themes provides a holistic understanding of how PDAs are used for people with depression in OMHCs in Saudi Arabia.

Emotional support

Four subthemes were identified from the analyses under the overarching main theme of emotional support:

- Encouragement
- Empathy
- Acknowledgment
- Offering kindness and well-wishing

The subthemes were analysed using sample data that mainly used PDAs such as likes, emojis and other visualization symbols.

Encouragement

In the sample, many users shared their achievements in their depression journey, while others encouraged them by using PDAs, for example:

> *by the grace of God, I've successfully overcome a depressive episode. And by God's will, others will be able to as well* ❤️🙏. (Samer, General)

Others responded using a variety of PDAs, as shown in Figure 3:



Figure 3. Reactions to Samer's post.

As a sense of achievement, Samer shared that he overcame a depression episode, and some users encouraged him by providing the above emojis, e.g., a thumbs-up sign to indicate approval, and an arm symbol to indicate strength. PDAs help people to offer support more easily and faster in a single click. This is especially useful in a depression context, where some people may have difficulty writing and reviewing the post. PDAs can facilitate the process of providing support to others. On the other hand, posters who receive these PDAs as a way to encourage them may be helped in terms of receiving various emojis, with different styles and frequencies for each PDA.

> *Just a day after my surrender, and collapse, I decided to return stronger. Of course, this strength comes from God, then your words* 🫂🫂🫂🫂🫂. (Thai, General)

Others responded with different emojis, as shown in Figure 4:

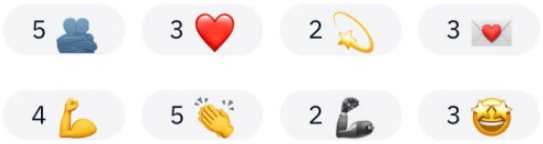


Figure 4. Reactions to Thai's post

Thai shared a thread that she got stronger and overcame her problems, while some users used PDAs to encourage her. The diverse PDAs used indicate various positive feelings for the poster, such as graduating, loving what she did, admiring her post, and applauding her. All these affirmative feelings can provide emotional support and encouragement to a person.

> *Today, I accomplished something really great. And I have no one to share my joy with, so please cheer me on!* (Hessah, General)

Others responded, as shown in Figure 5:



Figure 5. Responses to Hessah's post

Hessah felt loneliness and wanted to share her achievements with others. While some users used PDAs to offer support, she received diverse PDAs that can be interpreted with various positive feelings, such as encouragement and love.

The above quotations demonstrate how Samer, Thai, and Hessah's posts elicited affirmative reactions. This highlights the pivotal role of PDAs in efficiently enhancing the process of offering support. Additionally, people who read those threads and view the number of PDAs received can be motivated to do the same.

Empathy

Several posts offered empathy to others who are suffering and facing difficulties caused by depression. Many posters stated that they felt lonely and isolated from the physical world. Online communities can offer them love, a sense of belonging, and empathy, which they may not be receiving from their families or society due to social and cultural factors. For example, expressing

problems or sharing emotions with others can be considered a weakness. (Even when people offer emotional support or love, they may view vulnerability as a sign of weakness). Utilizing nonverbal PDAs can enhance the ability to express empathy and understanding toward others. For example,

> *this is my first time writing something here. I feel this tightness, like I've lost everything good, and the stress, lack of sleep, and negative thoughts are increasing every day. I've become someone unlikeable, lonely, and depressed. I feel like people planted these traits in me, or maybe I have been like this all along. I feel lost and like I have no support … The proof of the loneliness I've reached is that I couldn't find anyone to say this to, so I wrote it for strangers. I wished a stranger could understand me* (Abdul-Karim, General).

Zyana responded,

> *I feel you* 🫂😭 . (Zyana, General)

Abdul-Karim expressed their struggle through their depression journey and the loneliness they felt. Using online communities helped them to share their feelings and seek the support they needed. Others, such as Zyana, responded to Abdul-Karim by offering support and empathy. Zyana felt Abdul-Karim's loneliness and empathized with him. Using PDAs in this context can help to easily interpret the emotions Zyana felt for Abdul-Karim. Although the emoji used (crying) here can be interpreted as a negative emotion, receiving such an emoji in this context may indicate that users feel or understand the poster's emotions and pain, which can be interpreted as comforting positive feelings.

> *Unemployment destroys a person in many ways … I feel hopeless … I feel like I have no value, and life without a job feels depressing and meaningless.* (Reem, General)

Others responded, as shown in Figure 6:



Figure 6. Responses to Reem's post

Reem's depression was partly caused by unemployment. She shared her story with others to express her current feelings and, in response, people used various PDAs to show support. These gestures could be interpreted as empathy and sympathy, offering positive emotional reinforcement. The use of PDAs not only helps the recipient feel supported but can also benefit the sender. By simply selecting an emoji that conveys their emotions, the sender may experience a sense of satisfaction and pleasure from helping others.

Acknowledgment

Various people who posted reassured others with their opinions, which may indicate an understanding of the person's situation:

> *some people don't know how to talk about their pain. If they want to tell you they're hurting, they bite their nails, turn their face away from you, and embrace silence … If they sigh, they sigh deeply; if they cry, they cry in silence. When they want to express their feelings, it manifests as a lump in their throat. Do you understand what it means when nothing senses or understands your pain, except the pillow?* (Nawal, General)

Nawal posted a common issue, which is that people can hide their depression and try to avoid the situation in the physical world. In Saudi Arabia, some people do not accept depression as a real condition; this may be due to cultural traditions or wrong beliefs. Thus, people hide their depression and pretend that nothing is happening. Online communities offer a free place to access and share stories with people in similar situations. Using various PDAs, many people agreed with what Nawal posted, as indicated in some noting their full acknowledgment.

A further post, by Saafe below, provided advice for others:

> *good morning. Consistency is the foundation of treatment. In my opinion, the doctor gives you a treatment plan and stays with you throughout the process. That's why it's important to stick with your doctor, and believe me, you will become stronger and a valuable member of society.* (Saafe, Self-Development)

Others responded, as shown in Figure 7:



Figure 7. Responses to Saafe's post

Saafe shared his opinions, which were based on his experience and offered a sense of hope while others acknowledged his views. Offering an agreed opinion using emojis as well as the number of people who agree can further encourage users to engage more actively in online communities.

Offering kindness and well-wishing

When members express their isolation or other bad feelings, others often respond with kindness and well wishes as a form of emotional support and include PDAs themselves.

> *I wish you a happy year in all aspects and better mental health* 🥳🎂🎈🎊🎉. (Amal, General)

Amal replied to a poster stating that it was her birthday, and she was lonely, with no one celebrating with her. Other posters replied to her, wishing her a happy year, just like Amal. The types of PDAs used indicated that Amal was very happy, as there was a virtual party with a cake and a balloon suggesting a cause for celebration. These emojis can help with receiving positive emotions and reducing isolation.

In contrast, Naasira expressed she felt exhausted and in pain:

> *Oh God, I'm tired.* (Naasira, General)

For which others offered her kindness and wished her well, as shown in Figure 8:



Figure 8. Responses to Naasira's post.

Using PDAs such as the plaster emoji can be an indicator of wishing for better mental health as a sense of wishing for recovery, and others replied by praying. These affordances enable people with depression in online communities to access, share, and facilitate communications that foster

emotional support with a set of various PDAs, and more can help in engaging in online communities.

Informational support

Many posters used PDAs to reduce ambiguity and facilitate the interpretation of information. Additionally, PDAs can help clarify intentions and perceptions.

> *Always keep a (shield of protection* 🛡️*) from people's negative words, whatever the type. We all get affected by words, especially negative ones, which is very normal. But within limits. Don't give anyone the right to hurt you, making you reach a point where you start questioning yourself: 'Should I forgive them or not?'* (Sherine, General)

Sherine is an example of a poster who shared strategies for addressing negative thoughts and feelings. Sherine used a picture of a shield to help process the information and make the point clearer, so those who read this post can better understand her meaning, facilitating communication between them. Using PDAs can reduce the potential for misinterpretation and misunderstanding.

Peers of online support groups suggested trying new activities to discover enjoyable experiences that can lift their mood and bring a sense of pleasure.

> *A New experience, a new feeling* ✨🥂.
>
> *Advice: Try anything new to your food, places, hobbies, and styles and experience a new feeling.* (Malak, General)

Other posters responded using a variety of PDAs, as shown in Figure 9:

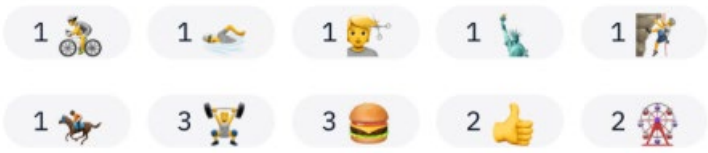


Figure 9. Responses to Malak's post.

Malak initiated a thread and suggested that others try new things and explore their feelings. Some posters responded with emojis representing places or activities they had tried or wanted to try. By viewing the diverse images used in this thread, people may be motivated to try new things, imagining activities and places that can offer them a sense of pleasure and enhance their mood.

> *Good morning. I still suffer from sleep disturbances, but despite all of that,*
>
> *I studied and completed a small section ✓.*
>
> *I went to the gym ✓.*
>
> *I prayed in the mosque ✓.*
>
> *I made my food by myself, and I ate somewhat healthy food ✓.*
>
> *I'm trying to stay active and live my life, even if I have some difficulties.* (Mashal, General)

Meshal highlighted that he does these daily tasks with a sense of achievement, which could motivate others to do the same. Using a nonverbal symbol offers the poster a sense of accomplishment and relief and, therefore, a better mood, especially the feeling when he tagged the activities. Additionally, such a symbol could have helped receivers with a quick and easy interpretation of the post.

> *Suicidal thoughts and self-harm attempts + self-criticism + guilt + distress and sadness … and so much more* 🔴*. I really wish someone could give me a solution, treatment, or help for these things I'm struggling with. I'm literally exhausted beyond what you can imagine.* 🔴 *Please help me—my mental health is deteriorating, and my spirit is worn out* 🆘. (Rani, General)

Rani expressed that she was seriously contemplating suicide, and she used PDAs to share the need for help. The use of red flags as signs for emergencies helps to direct attention to what has been shared. These alert symbols can add an emotional tone to the text, highlighting the seriousness of the message, which then accelerates replies offering help by sharing the information needed. This suggests that PDAs play an important role in receiving informational support.

PDAs as a harmful tool

Despite the important role of PDAs in receiving emotional and informational support, PDAs can be used to convey frustrating emotions. When users in online communities use negative emojis and add them in a context in which people expect to receive positive feelings, such emotions can be harmful, especially for people with depression. Receiving frustrated emojis can lead to being less active in online communities, which can increase loneliness and isolation.

> *Peace be upon you (Salaam alaikum). Where have you been? Everyone asked about you. Come back strong and don't let anyone affect you. Is it because of the bone emoji? They will get bored and leave you. Be strong.* (Khaled, General)

Khaled posted to Nouf, who had previously started a thread and received a bone emoji, which symbolizes negativity. Nouf was active in the forum, but later, she became less active.

Discussion

This study has highlighted the role of PDAs in offering emotional and informational support while also revealing their potential to cause harm for people with depression in OMHCs. Using posts that include text only or both text and emojis can lead to emotional support (Benson et al., 2020; Mejova & Hommadova Lu, 2022). Encouragement, empathy, acknowledgment, kindness, and well-wishing are forms of emotional support. In OMHCs, people with depression can communicate and understand their emotions and feelings. By offering empathy and encouragement on these platforms, isolation can be reduced. Thus, the feeling of receiving emotional support can enhance engagement in such online communities, a finding consistent with those of previous studies on online health communities (e.g., Liu et al., 2020).

People with depression use PDAs to encourage others, as noted by (Baek et al., 2022; Choi et al., 2023), which are studies in an organizational context. Baek et al. (2022) suggested that brands use emojis to encourage customers to engage more on social media, while Choi et al. (2023) indicated that managers use positive emojis to encourage employees' creativity.

The use of PDAs in OMHCs to express understanding and empathy is similar to the findings of (Benson et al., 2020; Johnson et al., 2020; Mejova & Hommadova Lu, 2022). Benson et al. (2020) investigated the role of online health communities for informal caregivers of patients with cancer in hospice care. The study indicated that using emojis with sad expressions in posts indicates empathy and the disclosure of feelings. Mejova et al. (2022) observed a set of emojis inducing hugs

on Twitter, highlighting that emojis can be used to show care and empathy on social media. Additionally, Johnson et al. (2020) investigated posts with text and emojis on Instagram and stated that they can be interpreted as empathy, a form of emotional support. While these studies are considered important, they did not examine OMHCs.

PDAs can be used in OMHCs to convey acknowledgment as well as offer well-wishes and kindness. Johnson et al (2020) investigated that emojis can be an indicator of well-wishing, a form of positive emotional support. Furthermore, some emojis are used for reassuring (Dos Reis et al., 2016; Smith et al., 2023).

The current study confirmed an important aspect of PDAs, which is that they can foster informational support. Similarly, Barach et al. (2021) outlined how emojis serve multiple functions that enhance reading comprehension. Eye tracking was used to investigate the timing of semantic processing of emojis while reading. The inclusion of an emoji, compared to its absence, led to higher rates of skipping the target word and decreased the total time spent on those words. Additionally, Dos Reis et al. (2016) indicated that, in contrast to text-based communication, incorporating emojis and icons, such as emotional expressions and gestures, improves the quality of information.

Despite the important role of using PDAs in fostering social support, they can be used in a harmful way in OMHCs, as highlighted by Kusaila and Gerhart (2023), who indicated that the use of emojis in an organizational context can create uncertainty and, in certain cases, may result in frustration.

Conclusion

This study provides a perspective on PDAs used in online mental health contexts and is the first study of its kind on the use of OMHCs by people with depression in Saudi Arabia. The findings make several contributions. The study used thematic analysis to investigate the role of PDAs for people with depression in OMHCs and enhances our understanding of how people use these in this context. Second, using PDAs can assist in recognizing emotional and informational support and encouraging people with depression to become more engaged in online communities, which can help foster social support and reduce feelings of isolation.

Additionally, it may be used as a harmful tool, mainly when people receive frustrating emojis. These findings can offer guidance for designing OMHCs for people with depression to maximize the benefits and reduce the risks associated with using OMHCs by enhancing the alignment between the needs of people with depression and the ability of their support networks to address those needs. Future studies can benefit from the research direction that we provided.

Future research could involve conducting interviews with users with depression to investigate and explore the subsequent changes, such as cognitive and emotional ones, after using or receiving PDAs. Additionally, interviewing users with depression can offer their feedback on PDA use, which helps in improving the use of positive PDAs and reducing the negative ones.

This study is not without limitations since it employed qualitative methods, which are not generalizable to a broader population. Furthermore, the study collected posts from Saudi OMHCs but did not investigate the role of PDAs in other countries. However, although the findings may not be generalizable, they may be transferable to people with other mental health conditions using OMHCs and in other countries, for example, in the Middle East. Future studies could verify this and investigate the role of PDAs in other chronic diseases, in different countries, or even on different social media platforms.

Acknowledgements

The research was funded through scholarship provided by the Saudi government and the Ministry of Education in Saudi Arabia.

About the authors

Sara Alrashed is a PhD student in the Information School at University of Sheffield, Sheffield UK. Sara's interests include online mental health communities and how people use technology to share health Information.

Peter A. Bath is Professor of Health Informatics in the Information School, University of Sheffield, Sheffield, UK. He received his Ph.D. from the University of Sheffield, and his research interests include the use of health information by patients and the public and how people share information, emotions, and experiences in online health communities.

Kushwanth Koya is Kushwanth Koya is a Lecturer in Information Management in the Information School, University of Sheffield, Sheffield, UK. He received his PhD from Northumbria University, Newcastle-upon-Tyne and his research interests are at the intersection of society, information needs and digital technologies, specifically investigating how different sections of society have their information needs met through accessing various digital technologies.

References

Afful-Dadzie, E., Afful-Dadzie, A., & Egala, S. B. (2023). Social media in health communication: A literature review of information quality. *Health Information Management Journal*, 52(1), 3–17. https://doi.org/10.1177/1833358321992683

Alhabeeb, A. A., Al-Duraihem, R. A., Alasmary, S., Alkhamaali, Z., Althumiri, N. A., & BinDhim, N. F. (2023). National screening for anxiety and depression in Saudi Arabia 2022. *Frontiers in Public Health*, 11, 1–12. https://doi.org/10.3389/fpubh.2023.1213851

Argyris, Y. A., & Monu, K. (2015). Corporate use of social media: Technology affordance and external stakeholder relations. *Journal of Organizational Computing and Electronic Commerce*, 25(2), 140–168. https://doi.org/10.1080/10919392.2015.1033940

Baek, T. H., Kim, S., Yoon, S., Choi, Y. K., Choi, D., & Bang, H. (2022). Emojis and assertive environmental messages in social media campaigns. *Internet Research*, 32(3), 988–1002. https://doi.org/10.1108/INTR-11-2020-0638

Barach, E., Feldman, L. B., & Sheridan, H. (2021). Are emojis processed like words?: Eye movements reveal the time course of semantic processing for emojified text. *Psychonomic Bulletin and Review*, 28(3), 978–991. https://doi.org/10.3758/s13423-020-01864-y

Benson, J. J., Oliver, D. P., Washington, K. T., Rolbiecki, A. J., Lombardo, C. B., Garza, J. E., & Demiris, G. (2020). Online social support groups for informal caregivers of hospice patients with cancer. *European Journal of Oncology Nursing*, 44, 101698. https://doi.org/10.1016/j.ejon.2019.101698

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Caliandro, A., & Anselmi, G. (2021). Affordances-based brand relations: An inquire on memetic brands on Instagram. *Social Media+ Society*, 7(2). https://doi.org/10.1177/2056305121102136

Carr, C. T., Wohn, D. Y., & Hayes, R. A. (2016). As social support: Relational closeness, automaticity, and interpreting social support from paralinguistic digital affordances in social media. *Computers in Human Behavior*, 62, 385–393. https://doi.org/10.1016/j.chb.2016.03.087

Choi, J., Shim, S. H., & Kim, S. (2023). The power of emojis: The impact of a leader's use of positive emojis on members' creativity during computer-mediated communications. *Plos One*, *18*, 1–18. https://doi.org/10.1371/journal.pone.0285368

Clarke, V., & Braun, V. (2017). Thematic analysis. *The Journal of Positive Psychology*, *12*(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613

Cobb, S. (1976). Social support as a moderator of life stress. *Psychosomatic Medicine*, *38*(5), 300–314. https://doi.org/10.1097/00006842-197609000-00003

Cohen, S., & Hoberman, H. M. (1983). Positive events and social supports as buffers of life change stress. *Journal of Applied Social Psychology*, *13*(2), 99–125. https://doi.org/10.1111/j.1559-1816.1983.tb02325.x

Conte, R., & Paolucci, M. (2001). Intelligent social learning. *Journal of Artificial Societies and Social Simulation*, *4*(1), U61–U82.

Dos Reis, J. C., Jensen, C., Bonacin, R., Hornung, H., & Baranauskas, M. C. C. (2016). Expressive icons for the communication of intentions. *ICEIS 2016 - Proceedings of the 18th International Conference on Enterprise Information Systems*, *2*, 388–399. https://doi.org/10.5220/0005834003880399

Durbano, F., Irtelli, F., & Marchesi, B. (2024). *Depression - What Is New and What Is Old in Human Existence* (F. Durbano, F. Irtelli, & B. Marchesi (eds.); (Frist edition). https://doi.org/10.5772/intechopen.1000384

Gibson, J. J. (1977). The concept of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (pp. 67–82). Johne Wiley & Sons. http://macs-eu.org/slides/02_MACS_Affordances s-o-a_METU.pdf

Gibson, J. J. (2014). *The ecological approach to visual perception* (classic edition). Psychology press.

Gu, D., Li, M., Yang, X., Gu, Y., Zhao, Y., Liang, C., & Liu, H. (2023). An analysis of cognitive change in online mental health communities: A textual data analysis based on post replies of support seekers. *Information Processing and Management*, *60*(2), 103192. https://doi.org/10.1016/j.ipm.2022.103192

Hargreaves, S., Bath, P. A., Duffin, S., & Ellis, J. (2018). Sharing and empathy in digital spaces: Qualitative study of online health forums for breast cancer and motor neuron disease (amyotrophic lateral sclerosis). *Journal of Medical Internet Research*, *20*(6), e222. https://doi.org/10.2196/jmir.9709

Hasan, M. M., Al Amin, M., Arefin, M. S., & Mostafa, T. (2023). Green consumers' behavioral intention and loyalty to use mobile organic food delivery applications: the role of social supports, sustainability perceptions, and religious consciousness. *Environment, Development and Sustainability*, 1–51. https://doi.org/10.1007/s10668-023-03284-z

Hayes, R. A., Carr, C. T., & Wohn, D. Y. (2016a). It's the audience: Differences in social support across social media. *Social Media+ Society*, *2*(4), 1–12. https://doi.org/10.1177/2056305116678894

Hayes, R. A., Carr, C. T., & Wohn, D. Y. (2016b). One Click, Many Meanings: Interpreting Paralinguistic Digital Affordances in Social Media. *Journal of Broadcasting and Electronic Media*, *60*(1), 171–187. https://doi.org/10.1080/08838151.2015.1127248

Islam, T., Sheikh, Z., Hameed, Z., Khan, I. U., & Azam, R. I. (2018). Social comparison, materialism, and compulsive buying based on stimulus-response-model: a comparative study among adolescents and young adults. *Young Consumers*, *19*(1), 19–37. https://doi.org/10.1108/YC-07-

2017-00713

Johnson, B. L., Quinlan, M., & Pope, N. (2020). “Sticky Baby Dust” and Emoji: Social Support on Instagram during In Vitro Fertilization. *Rhetoric of Health & Medicine*, 3(3), 320–349. https://doi.org/10.5744/rhm.2020.1017

Johnson, D. S., & Lowe, B. (2015). Emotional Support, Perceived Corporate Ownership and Skepticism toward Out-groups in Virtual Communities. *Journal of Interactive Marketing*, 29(1), 1–10. https://doi.org/10.1016/j.intmar.2014.07.002

Kanchan, S., & Gaidhane, A. (2023). Social Media Role and Its Impact on Public Health: A Narrative Review. *Cureus*, 15, 1–10. https://doi.org/10.7759/cureus.33737

Karahanna, E., Xu, S. X., Xu, Y., & Zhang, N. A. (2018). The needs–affordances–features perspective for the use of social media. *Mis Quarterly*, 42(3), 737–756. https://www.jstor.org/stable/26635051

Krause, N. (1986). Social Support, Stress, and Well-Being Among Older Adults. *Journal of Gerontology*, 41(4), 512–519. https://doi.org/10.1093/geronj/41.4.512

Kusaila, D., & Gerhart, N. (2023). Sarcastic emojis: why don’t you understand? *Global Knowledge, Memory and Communication*. https://doi.org/10.1108/GKMC-06-2023-0196

Leidner, D. E., Gonzalez, E., & Koch, H. (2018). An affordance perspective of enterprise social media and organizational socialization. *The Journal of Strategic Information Systems*, 27(2), 117–138. https://doi.org/10.1016/j.jsis.2018.03.003

Lin, X., & Kishore, R. (2021). Social media-enabled healthcare: A conceptual model of social media affordances, online social support, and health behaviors and outcomes. *Technological Forecasting and Social Change*, 166, 120574. https://doi.org/10.1016/j.techfore.2021.120574

Liu, S., Xiao, W., Fang, C., Zhang, X., & Lin, J. (2020). Social support, belongingness, and value co-creation behaviors in online health communities. *Telematics and Informatics*, 50, 101398. https://doi.org/10.1016/j.tele.2020.101398

Lopez, A. D., & Murray, C. C. (1998). The global burden of disease, 1990–2020. *Nature Medicine*, 4(11), 1241–1243. https://doi.org/10.1038/3218

Lovatt, M., Bath, P. A., & Ellis, J. (2017). Development of trust in an online breast cancer forum: a qualitative study. *Journal of Medical Internet Research*, 19(5), e175. https://doi.org/10.2196/jmir.7471

Lu, J., Wang, X., Chen, S., Chen, G., Feng, Y., & Liu, L. (2024). Leveraging digital technology to improve self-efficacy in response to public health crises. *Information and Management*, 61(6), 103987. https://doi.org/10.1016/j.im.2024.103987

Lyubenko, R. (2023). Pros and Cons of Online Social Support Exchange on Social Networking Sites: a User’S Perspective. *Studia Universitatis Babes-Bolyai Sociologia*, 68(1), 103–125. https://doi.org/10.2478/subbs-2023-0004

Madhi, R. Bin, Khugali, A., & Kofi, M. (2020). Online health information seeking among patients in primary health care settings, Riyadh, Saudi Arabia. *International Journal of Advanced Community Medicine*, 3(3), 01–08. https://doi.org/10.33545/comed.2020.v3.i3a.151

Majchrzak, A., Faraj, S., Kane, G. C., & Azad, B. (2013). The Contradictory Influence of Social Media Affordances on Online Communal Knowledge Sharing. *Journal of Computer-Mediated Communication*, 19(1), 38–55. https://doi.org/10.1111/jcc4.12030

Marengo, D., Montag, C., Sindermann, C., Elhai, J. D., & Settanni, M. (2021). Examining the links between active Facebook use, received likes, self-esteem and happiness: A study using objective social media data. *Telematics and Informatics*, 58, 101523. https://doi.org/10.1016/j.tele.2020.101523

Mclaughlin, K. A. (2011). The public health impact of major depression: a call for interdisciplinary prevention efforts. *Prevention Science*, *12*, 361–371. https://doi.org/10.1007/s11121-011-0231-8

Mejova, Y., & Hommadova Lu, A. (2022). I feel you: Mixed-methods study of social support of loneliness on twitter. *Computers in Human Behavior*, *136*, 107389. https://doi.org/10.1016/j.chb.2022.107389

Nattembo, A. (2022). *Mediated Meaning: The Representation of Mental Health on Social Media Platforms* [Doctoral dissertation, The University of Adelaide]. https://hdl.handle.net/2440/137163

Norman, D. A. (1988). *The psychology of everyday things*. Basic books.

Nusawat, P., & Leelasantitham, A. (2024). A New Consistent Concept Between the I-change Model and the Transtheoretical Outcome to Imply the Leading Guideline of Depression Behavior Health Change Using Social Media. *The Open Public Health*, *17*(1), 1–20. https://doi.org/10.2174/0118749445318733240513061523

Odah, M., Ewis, A., Alkudaysi, F. M., Alhasani, T., Alessi, A. A., AlKudaysi, Y. M. Y., Alfaqih, K. A. M., Alfaqih, M. A. O., Almatrafi, A. A., Nmnkani, A. Z., Alzubaidi, A. S., Alothman, A. O., Alnashri, A. S., & Almathami, M. M. (2024). Prevalence of Depression Among the Adult Population in Southwestern Saudi Arabia: A Cross-Sectional, Community-Based Study. *Cureus*, *16*(3), e56081. https://doi.org/10.7759/cureus.56081

Page, X., Capener, A., Cullen, S., Wang, T., Garfield, M., & Wisniewski, P. J. (2022). Perceiving Affordances Differently: The Unintended Consequences When Young Autistic Adults Engage with Social Media. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3517596

Pfeifer, V. A., Armstrong, E. L., & Lai, V. T. (2022). Do all facial emojis communicate emotion? The impact of facial emojis on perceived sender emotion and text processing. *Computers in Human Behavior*, *126*, 107016. https://doi.org/10.1016/j.chb.2021.107016

Pope, C., & Mays, N. (2020). *Qualitative research in health care* (Fourth edition). Wiley Online Library.

Rawlings, L. (2023). *The Power of the "Like": A Quantitative Study on the Facebook Emoji as Social Support*. [Doctoral dissertation, Liberty University].

Ronzhyn, A., Cardenal, A. S., & Batlle Rubio, A. (2022). Defining affordances in social media research: A literature review. *New Media and Society*, *25*(11), 3165–3188. https://doi.org/10.1177/14614448221135187

Rostek, M. A., & Gamble, B. J. (2020). Origins: The community mental health and wellness futures forum. *Journal of Military, Veteran and Family Health*, *6*(3), 1–3. https://doi.org/10.3138/JMVFH-6.S3-0001

Shumaker, S. A., & Brownell, A. (1984). Toward a theory of social support: Closing conceptual gaps. *Journal of Social Issues*, *40*(4), 11–36. https://doi.org/10.1111/j.1540-4560.1984.tb01105.x

Smith, C. E., Miller Hillberg, H., & Levonian, Z. (2023). " Thoughts & Prayers" or "  &  ": How

the Release of New Reactions on CaringBridge Reshapes Supportive Communication in Health Crises. *Proceedings of the ACM on Human-Computer Interaction*, 7,1-39. https://doi.org/10.1145/3610035

Steinert, S., & Dennis, M. J. (2022). Emotions and Digital Well-Being: on Social Media's Emotional Affordances. *Philosophy and Technology*, 35(2), 1–21. https://doi.org/10.1007/s13347-022-00530-6

Sun, S., Seo, M., Wang, F., & Liu, Z. (2021). Social Support and Connective Affordances: Examining Responses to Early Covid-19 Patient Support Seeking on Microblogs. *Journal of Broadcasting and Electronic Media*, 65(5), 621–640. https://doi.org/10.1080/08838151.2021.2011284

Swain, J. (2018). A hybrid approach to thematic analysis in qualitative research: Using a practical example. *Sage Research Methods*, 10. https://doi.org/10.4135/9781526435477

Tang, J. L. (2022). Are You Getting Likes as Anticipated? Untangling the Relationship between Received Likes, Social Support from Friends, and Mental Health via Expectancy Violation Theory. *Journal of Broadcasting and Electronic Media*, 66(2), 340–360. https://doi.org/10.1080/08838151.2022.2087654

The University of Sheffield. (2023). *Ethics Policy Governing Research Involving Human Participants, Personal Data and Human Tissue: General Principles and Statements*. The University of Sheffield. https://students.sheffield.ac.uk/research-ethics/ethics/updates

Thoits, P. A. (1982). Conceptual, methodological, and theoretical problems in studying social support as a buffer against life stress. *Journal of Health and Social Behavior*, 23(2), 145–159. https://doi.org/10.2307/2136511

Treem, J. W., & Leonardi, P. M. (2013). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association*, 36(1), 143–189. http://dx.doi.org/10.2139/ssrn.2129853

Vaast, E., Safadi, H., Lapointe, L., & Negoita, B. (2017). Social media affordances for connective action: An examination of microblogging use during the Gulf of Mexico oil spill. *MIS Quarterly*, 41(4), 1179–1205. https://www.jstor.org/stable/26630290

World Health Organization(WHO). (2023, March 31). *Depressive disorder (depression)*. https://www.who.int/news-room/fact-sheets/detail/depression

Wu, M., Wu, T., & Pei, Y. (2024). What Drives Health Information Exchange on Social Media? Social Media Affordances and Social Support Perspectives. *Health Communication*, 1–15. https://doi.org/10.1080/10410236.2024.2321408

Xiao, W., Peng, J., & Liao, S. (2022). Exploring the associations between social media addiction and depression: Attentional bias as a mediator and socio-emotional competence as a moderator. *International Journal of Environmental Research and Public Health*, 19(20), 13496. https://doi.org/10.3390/ijerph192013496

Xu, K., & Li, X. (2023). Complaining, Regret, Superiority, and Discovery: Chinese Patients' Sense Making of Depression in an Online Forum. *Qualitative Health Research*, 33(7), 613–623. https://doi.org/10.1177/10497323231160119

Yoon, J. W., & Syn, S. Y. (2023). The effects of visual formats on Facebook health-related posts: evidence from eye movement analysis. *Library Hi Tech*, 41(3), 833–852. https://doi.org/10.1108/LHT-10-2021-0360

Zeitouni, N., & Milstein, I. (2017). The role of social support in Teachers' intention to use ICT tools

in their classrooms. 2017 *Portland International Conference on Management of Engineering and Technology* (PICMET), 1–5. https://doi.org/10.23919/PICMET.2017.8125274

Zhang, X., Liu, S., Chen, X., Wang, L., Gao, B., & Zhu, Q. (2018). Health information privacy concerns, antecedents, and information disclosure intention in online health communities. *Information & Management*, 55(4), 482–493. https://doi.org/10.1016/j.im.2017.11.003

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Enhancing vaccine communication in social Q&A: identifying readily applicable factors for answer acceptance on medical sciences stack exchange

Hengyi Fu

DOI: https://doi.org/10.47989/ir30iConf47155

Abstract

Introduction. Despite the efficacy of vaccines, online misinformation contributes to vaccine hesitancy. This study investigates factors influencing the acceptance of answers to vaccine-related questions on social Q&A platforms, aiming to improve online vaccine communication.

Method. We analysed 511 vaccine-related questions and 875 associated answers on the medical sciences stack exchange platform. Eleven variables were identified as potential predictors of answer acceptance, including citation practices, use of personal experiences, and recognition of asker's concerns.

Analysis. Logistic regression was employed to determine which variables significantly predict the likelihood of an answer being designated as the '*accepted answer*' by the asker.

Results. Personal experience sharing, recognition of asker's concerns, citing data sources, and including calls-to-action emerged as significant predictors of answer acceptance. Notably, professional credentials had a modest effect, while factors such as the use of analogies or citing health authorities were not significant predictors.

Conclusions. Our findings suggest that effective vaccine communication on social Q&A platforms should balance factual information with personal narratives, address user concerns empathetically, and provide clear, actionable advice. These insights can inform strategies for health communicators and platform designers to enhance the quality and acceptance of vaccine information online.

Introduction

Vaccines have prevented millions of deaths and substantially reduced the burden of infectious diseases globally. However, despite their proven efficacy, vaccine hesitancy has emerged as a significant global health concern. The World Health Organization has identified vaccine hesitancy as one of the top ten threats to global health, highlighting the urgent need for effective vaccine communication strategies (World Health Organization, 2019).

The advent of social media has dramatically altered the landscape of health information dissemination, particularly regarding vaccines. Social Q&A communities have evolved into significant platforms where users seek and share advice, knowledge, and experiences on various health topics, including vaccination (Sharon et al., 2020). The exchange of vaccination-related questions and answers has become particularly prominent during the COVID-19 pandemic, reflecting increased concern for understanding vaccination protocols and their effects (Wang et al., 2022, Fu and Oh, 2023). In this context, the ability to craft convincing and accurate responses to vaccine-related questions on social Q&A platforms has become crucial. Health professionals, public health communicators, and concerned citizens engaging in these spaces face the challenge of not only providing factual information but also addressing emotional concerns, countering misinformation, and fostering vaccine confidence (Chou and Budenz, 2020).

Thus, the fundamental question this project aims to answer is: how can we provide convincing and persuasive responses to online vaccine-related information seeking? Our focus is on identifying factors that are easily recognized and implemented by general users when writing responses on social Q&A platforms and other online forums, and also can be easily incorporated into community's policy system or health education materials.

In this paper, we report our preliminary findings on accepted answers selected by askers for vaccine-related questions on the medical science stack exchange Q&A site. Specifically, we explore what variables can predict whether askers will designate a given answer to a vaccine-related question as the '*accepted answer*.' Unlike features such as readability score, which can be difficult for ordinary users to consistently improve, we focus on factors that answerers can easily be conscious of and implement, and that can be readily applied by health professionals, online community designers, and concerned citizens engaging in online vaccine discussions (Oh et al., 2012, Kienhues et al., 2011).

By focusing on these readily applicable factors, we aim to provide practical insights that can be put into action by individuals seeking to contribute positively to online vaccine discussions. This research not only contributes to improving public health communication strategies but also empowers individuals to play an active role in combating misinformation and promoting vaccine confidence in digital spaces (Vraga and Bode, 2017).

It's important to interpret the results of this paper carefully. Right now, our study doesn't suggest a cause-effect relationship. We're not saying that the variables/factors identified are the only or the full set behind the difference of accepted or non-accepted answers.

Previous studies

Despite the proven efficacy of vaccines in preventing disease and saving lives, vaccine hesitancy has emerged as a significant concern, leading to delays and refusals in vaccination. This scepticism is often combined with mistrust and suspicion of health providers and medical companies, potentially compromising their impartiality (Sharon et al., 2020). During the COVID-19 pandemic, such dynamics have intensified, with misinformation significantly influencing vaccination decisions.

Research has shown that users evaluate answers based on multiple criteria, including completeness of information, solution feasibility, and the answerer's attitude (Kim et al., 2009), with answer quality influenced by the answerer's reputation and supporting references (Shah & Pomerantz, 2010). Harper et al. (2008) found that answer length, references, and the answerer's history were significant predictors of answer acceptance. Zhang (2014) found that users value both expertise indicators and relatable personal experiences when evaluating health-related answers. These findings complement Moran et al.'s (2016) analysis of persuasive techniques in anti-vaccine websites, particularly noting the impact of personal narratives and scientific evidence.

Kienhues et al. (2011) highlighted the effectiveness of balanced information presentation, including potential risks and benefits, which informed our examination of factors like recognizing asker's concern and citing multiple sources. This aligns with Vraga and Bode's (2017) findings that corrections were more effective when they explained why misinformation was incorrect rather than simply stating it was wrong.

Oh et al. (2016) found that users seeking health information value emotional support alongside factual information. Sharon et al. (2020) found that while few users explicitly sought expert testimony, answers from self-identified health professionals were twice as likely to be designated as '*best answers*.' Jang et al. (2022) and Xu and Guo (2018) demonstrated how emotional language and narrative approaches influenced vaccine information acceptance. Chou and Budenz (2020) emphasized that effective vaccine communication should address both factual information and emotional concerns, while Lunz Trujillo et al. (2021) found that stories highlighting non-vaccination consequences were particularly effective.

While these studies have addressed various factors influencing health information acceptance, no comprehensive model on predictive features for vaccine-related information has been tested in social Q&A platforms. Some predictive features, even when proven effective, are not easily applicable by ordinary users or community designers. Our study contributes by providing empirical evidence on factors influencing vaccine-related answer acceptance in a specialized Q&A community.

Method

Studying site and dataset

The medical sciences stack exchange (MSSE) (https://medicalsciences.stackexchange.com/) is a specialized online health Q&A platform within the broader Stack Exchange network, which hosts over 200 subject-oriented Q&A communities. Launched in 2015, MSSE has grown to become one of the most prominent platforms dedicated to health and medical sciences, with a community of 19,527 registered users as of January 2024. Unlike other Q&A platforms such as Yahoo! Answers or Quora, MSSE prioritizes community-driven development and self-governance, with members playing an active role in managing and moderating the platform. Additionally, MSSE actively encourages participation from health professionals to align with its mission of disseminating accurate medical knowledge. Another key advantage of MSSE is its commitment to data transparency and accessibility. The platform publishes all user-generated content, making it publicly available through its interface and application programming interface (API).

In this study, we extracted all questions posted on MSSE between March 31, 2015, and December 31, 2022, using the official API. The quantitative and interactive question features, including question score, view count, answers, answer count, and accepted answer status, were collected until March 31, 2023. A total of 6,760 questions and 6,505 answers were collected. Out of these, 511 vaccine-related questions were identified by retrieving questions containing the keywords '*vaccine(s)*', '*vaccination*' or related synonyms in the questions and were used for further analysis. There were in total 875 answers associated with the 511 vaccine-related questions. Among those answers, 323 answers (36.91%) were selected as accepted answers by the asker. A question can

have only one accepted answer selected by the asker, and some questions do not have any accepted answer. It is important to note that on MSSE, the acceptance of an answer is decided solely by the question asker. While this does not necessarily reflect a collective assessment of answer quality, it indicates whether the asker's information needs have been met.

Variable selection

The dependent variable represented whether the answer appeared as the '*accepted answer*' to the question. It was coded as 0 or 1 to signify '*no*' and '*yes*'. As a starting point, when choosing the independent variables, we began with those can be consciously controlled or easily incorporated into answers, including citing of research (Oh et al., 2012, Diviani et al., 2015), citing of data resources (Oh et al., 2012, Sbaffi and Rowley, 2017), citing health authorities (Vraga and Bode, 2017), citing other answers (Oh et al., 2012), use of hedging language (Mayweg-Paus and Jucks, 2015, Dunn et al., 2015), use of formatting for clarity (Meppelink et al., 2015, Lazard and Mackert, 2014), indicating him/her is a health professional (Hu and Shyam Sundar, 2010, Huh et al., 2013), including personal experience (Pian et al., 2020, Ziebland, 2012), including call-to-action (Ludolph et al., 2016, Korda and Itani, 2013), use of analogies and metaphors (Pelaccia et al., 2011, Reyna, 2012), recognizing asker's concern (even if the concern is not scientifically valid) (Goldstein et al., 2015, Amin et al., 2017). All these variables have been examined and proven effective in serving online health information needs in previous research. Table 1 lists explanations, measurements, and examples of each variable (We use a list of 111 terms of hedges in Bordignon et al. (2021)).

Variable	Description	Example	Value
if_accepted	If an answer is an accepted answer or not	N/A (determined by platform)	0 or 1
n_citing research	Number of research citations in the answer	'A study by ABC et al. (2022) found that vaccine efficacy remains high after six months…'	Integer ≥ 0
n_citing_datasource	Number of data sources cited in the answer	According to the CDC's COVID Data Tracker (https://covid.cdc.gov/covid-data-tracker), vaccination rates have increased by 15% since last month…"	Integer ≥ 0
n_citing_health_authorities	Number of times health authorities mentioned	'CDC recommended annual flu vaccines for everyone over 6 months old…'	Integer ≥ 0
n_citing_other_answer	Number of references to other answers	'To add to what @DrSmith said about vaccine safety in their answer above...' 'Building on the point about herd immunity raised by (answer link) in his response...'	Integer ≥ 0
n_hedging term/word	Number of hedging terms an answer includes	'It's ***possible*** that new variants could affect vaccine efficacy…'	Integer ≥ 0
if_formatting	If the answer uses formatting (bullet points or numbered lists)	'Common vaccine side effects:\n1. Soreness at injection site\n2. Mild fever\n3. Fatigue…'	0 or 1
if_health_pro	If the answerer indicates s/he is a health professional	'Speaking from my perspective as a family physician…'	0 or 1
if_personal_exprience	If the answer includes personal experience	'After getting my COVID-19 vaccine, I experienced mild arm soreness for also three days…'	0 or 1
if_call_to_action	If the answer includes call to action	'If you're concerned about vaccine side effects, I recommend keeping a symptom diary for a few days after receiving the vaccine…'	0 or 1
if_ analogies	If the answer uses analogies or metaphors	'The mRNA in vaccines is like a blueprint. It doesn't build anything itself…'	0 or 1
if_recognazing_concern	If the answer recognizes the asker's concern, even if it's not scientifically valid	'While I can assure you that vaccines do not contain microchips for tracking, your question reflects a broader concern about privacy and trust in medical interventions...'	0 or 1

Table 1. Variable description, example, and value

We employed a combination of automatic and manual coding methods to code variables for each answer. For the automatic coding process, we used Python to develop Regex (regular expression) rules for the initial screening of certain variables. These rules identified links or symbols associated with citing research, data resources, and other answers, as well as specific formatting elements such as bullet points and numbered lists. Additionally, Regex was utilized to detect keywords or phrases indicative of hedging language, such as '*possibly*' or '*it seems*' (Bordignon et al., 2021).

Manual coding was conducted to comprehensively annotate variables that required nuanced interpretation, such as recognizing the use of personal experiences or assessing whether an answer addressed the asker's concerns. This process involved two researchers who worked collaboratively to refine coding guidelines and resolve ambiguities in the annotation process. Both researchers independently coded a subset of the data to calculate intercoder reliability, achieving

a Cohen' s κ of 0.856, which indicates a high level of agreement. Discrepancies were resolved through discussion to ensure consistency across the dataset.

Automatic coding was also performed using the GNU '*Style and Diction*' software package (https://www.gnu.org/software/diction/) following Fu and Oh (2019) for more sophisticated writing style features, such as sentence complexity and vocabulary usage. However, these features were not included in our final analysis as they do not align with our focus on readily applicable factors.

Preliminary results and discussion

A logistic regression analysis was conducted using SPSS to examine variables influencing the likelihood of an answer being selected as the '*accepted answer*' in vaccine-related questions on Medical Science Stack Exchange. The model demonstrated moderate predictive power, with a Nagelkerke R^2 of 0.474, indicating that it explains 47.4% of the variance in accepted answer selection. The Omnibus Test of Model Coefficients indicated that the accuracy of the model exceeded that of the baseline model when adding the predictors (χ^2 (11) = 204.623, p < 0.001). The Hosmer & Lemeshow test of goodness of fit suggested the model was a good fit to the data (p = 0.279).

Variable	β	S.E. β	Wald's χ^2	df	p	e^β (odds ratio)
Constant	-1.885	0.145	168.905	1	<0.001 ***	0.152
if_personal_experience	0.391	0.079	24.503	1	<0.001***	1.478
if_recognizing_concern	0.287	0.082	12.252	1	<0.001***	1.332
n_citing_datasource	0.269	0.072	13.958	1	<0.001***	1.309
if_call_to_action	0.230	0.078	8.686	1	0.003**	1.259
n_citing_research	0.182	0.068	7.164	1	0.007**	1.200
if_health_pro	0.184	0.073	6.356	1	0.012*	1.202
n_citing_health_authorities	0.156	0.081	3.710	1	0.054	1.169
if_formatting	0.143	0.094	2.315	1	0.128	1.154
if_analogies	0.128	0.099	1.670	1	0.196	1.137
n_citing_other_answer	0.117	0.082	2.035	1	0.154	1.124
n_hedging_term	-0.074	0.051	2.105	1	0.147	0.929
Overall model evaluation			**χ^2**	**df**	**p**	
Omnibus Test of Model Coefficients			204.623	11	<0.001	
Goodness-of-fit Test			**χ^2**	**df**	**p**	
Hosmer & Lemeshow			9.8	8	0.279	
Nagelkerke R^2: 0.474						

Table 2. Logistic regression predicting the likelihood of selecting as '*accepted answer*'

The results show several factors that powerfully predict the likelihood of an answer being chosen as an accepted answer. Personal experience emerged as the strongest predictor (OR = 1.478, $p < 0.001$), suggesting that answers incorporating the writer's personal experiences are 47.8% more likely to be selected as accepted. This finding aligns with previous research that highlighted the value users place on personal experiences in online health information seeking. Recognizing the asker's concerns also proved to be a significant factor ($\beta = 0.287$, OR = 1.332, $p < 0.001$). While Medical Stack Exchange is a health Q&A community that emphasizes professional health information sharing and in general, does not recommend emotional support and too much personal experience (Fu and Oh, 2022), this highlights the importance of empathy and direct engagement with the asker's perspective in effective health communication.

The citation of data sources (OR = 1.309, $p < 0.001$) and research (OR = 1.200, $p = 0.007$) both significantly increased the odds of an answer being selected. This indicates that users value evidence-based responses, with each additional data source cited increasing the odds by 30.9% and each research citation increasing the odds by 20%. These findings emphasize the importance of backing claims with credible sources in vaccine-related discussions. These findings align with previous research emphasizing the importance of backing claims with credible sources in health-related online discussions.

Interestingly, including a call to action in the answer also positively influenced selection (OR = 1.259, $p = 0.003$). Answers that provided clear, actionable advice were 25.9% more likely to be chosen as best, suggesting that users appreciate practical guidance alongside information.

The identity of the answerer as a health professional had a modest effect (OR = 1.202, $p = 0.012$). Unlike other studies, this indicates that while professional credentials are valued, they are not the dominant factor in determining the perceived quality of an answer in this context.

Notably, several factors that might be expected to influence answer selection were found to be non-significant. These included citing health authorities, the use of formatting for clarity, employing analogies, referencing other answers, and the use of hedging language. The non-significance of these factors suggests that users may prioritize content and relevance over stylistic elements or appeals to authority. The non-significance of references to public health authorities may be caused by the distrust of public authorities during the pandemic but requires further investigation.

Implications, limitations, and future research

Our study identifies several key factors that predict the acceptance of vaccine-related answers on social Q&A platforms, which have implications for various stakeholders.

For ordinary users, implementing the identified factors can enhance the credibility and engagement of their responses. Our finding that personal anecdotes significantly increase the likelihood of answer acceptance empowers users to confidently share their own vaccine experiences. The fact that answer acceptance is not solely dependent on professional credentials but also on factors like citing data sources, encourages more users to contribute their knowledge. These insights are particularly valuable for users who might hesitate to contribute to vaccine discussions due to a perceived lack of professional credentials. By focusing on these accessible factors, ordinary users can craft more effective responses and play a more active role in online vaccine discussions.

While personal narratives can make information more relatable and engaging, they also present potential risks in health communication, particularly around vaccine information. Personal experiences might inadvertently spread misinformation if they present outlier cases as typical or contradict scientific evidence. One potential solution, supported by our findings, is to encourage the combination of personal experiences with data citations and research citations. This combined

approach could help maintain scientific accuracy while preserving the engaging aspects of personal narratives.

For platform designers, the insights from this study can directly inform the design of social Q&A platforms to facilitate more effective vaccine communication. Specific features could be integrated into the answer submission process, such as prompts encouraging users to share relevant personal experiences, built-in citation tools to easily reference credible sources, and suggestion systems that remind users to address the asker's specific concerns. Additionally, platforms could refine their algorithms to prioritize answers exhibiting these characteristics. For example, answers that include data citations and address user concerns could be given higher visibility, potentially improving the overall quality of information presented to users.

Health communicators can tailor their messages by incorporating those factors to increase relatability, consistently citing reputable data sources to enhance credibility. These strategies can be incorporated into health communication curricula, with specific modules on effective online communication. For instance, training exercises could focus on crafting responses that balance personal experience, scientific evidence, and empathetic addressing of concerns. This approach could lead to a new generation of health communicators better equipped to engage in digital health discussions, particularly around vaccines.

This study has several limitations. Firstly, our analysis is based on data from a single platform, Medical Science Stack Exchange, which may not be representative of all social Q&A environments or user demographics. Secondly, the study focuses on a specific time period, which may not capture long-term trends or changes in user behavior, especially given the rapidly evolving nature of vaccine discourse during the COVID-19 pandemic. Future research could address these limitations by expanding the analysis to multiple platforms and conducting longitudinal studies to track changes over time. Additionally, qualitative research methods, such as interviews or focus groups with users, could provide deeper insights into the reasoning behind answer acceptance and the nuanced ways in which users interpret and value different answer characteristics.

About the author

Hengyi is an Assistant Professor in the School of Library and Information Studies, at the University of Alabama. Her primary research interests are online peer production, health information behaviour/interaction, and AI implementation in libraries.

References

Amin, A. B., Bednarczyk, R. A., Ray, C. E., Melchiori, K. J., Graham, J., Huntsinger, J. R., & Omer, S. B. (2017). Association of moral values with vaccine hesitancy. Nature Human Behaviour, 1(12), 873-880.

Bordignon, F., Ermakova, L., & Noel, M. (2021). Over-promotion and caution in abstracts of preprints during the COVID-19 crisis. Learned Publishing, 34(4), 622-636. Chou, W.-Y. S., & Budenz, A. (2020). Considering emotion in COVID-19 vaccine communication: addressing vaccine hesitancy and fostering vaccine confidence. Health communication, 35(14), 1718-1722.

Diviani, N., Van Den Putte, B., Giani, S., & van Weert, J. C. (2015). Low health literacy and evaluation of online health information: a systematic review of the literature. Journal of Medical Internet Research, 17(5), e112.

Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D., & Coiera, E. (2015). Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. Journal of Medical Internet Research, 17(6), e4343.

Fu, H., & Oh, S. (2019). Quality assessment of answers with user-identified criteria and data-driven features in social Q&A. Information Processing and Management, 56(1), 14–28. https://doi.org/https://doi.org/10.1016/j.ipm.2018.08.007

Fu, H., & Oh, S. (2022). Online community development in the early stages: the life cycle model application to Medical Sciences Stack Exchange. Aslib Journal of Information Management, 74(6), 1214-1232. https://doi.org/10.1108/AJIM-11-2021-0345

Fu, H., & Oh, S. (2023). Topics of questions and community interaction in social Q&A during the COVID-19 pandemic. Health Information & Libraries Journal. https://doi.org/10.1111/hir.12506

Goldstein, S., MacDonald, N. E., & Guirguis, S. (2015). Health communication and vaccine hesitancy. Vaccine, 33(34), 4212-4214. https://doi.org/https://doi.org/10.1016/j.vaccine.2015.04.042

Hu, Y., & Shyam Sundar, S. (2010). Effects of online health sources on credibility and behavioral intentions. Communication research, 37(1), 105-132.

Huh, J., Yetisgen-Yildiz, M., & Pratt, W. (2013). Text classification for assisting moderators in online health communities. Journal of biomedical informatics, 46(6), 998-1005.

Jang, S. H., Gerend, M. A., Youm, S., & Yi, Y. J. (2022). Understanding coronavirus disease 2019 (COVID-19) vaccine hesitancy: Evidence from the community-driven knowledge site Quora. Digital Health, 8, 20552076221145426. https://doi.org/10.1177/20552076221145426

Kienhues, D., Stadtler, M., & Bromme, R. (2011). Dealing with conflicting or consistent medical information on the web: When expert information breeds laypersons' doubts about experts. Learning and Instruction, 21(2), 193-204.

Kim, S., Oh, S., & Oh, J. S. (2009). Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective. Proceedings of the American Society for Information Science and Technology, 45(1), 1-15.

Korda, H., & Itani, Z. (2013). Harnessing social media for health promotion and behavior change. Health promotion practice, 14(1), 15-23.

Lazard, A., & Mackert, M. (2014). User evaluations of design complexity: The impact of visual perceptions for effective online health communication. International journal of medical informatics, 83(10), 726-735.

Ludolph, R., Allam, A., & Schulz, P. J. (2016). Manipulating Google's knowledge graph box to counter biased information processing during an online search on vaccination: Application of a technological debiasing strategy. Journal of Medical Internet Research, 18(6), e137.

Lunz Trujillo, K., Motta, M., Callaghan, T., & Sylvester, S. (2021). Correcting misperceptions about the MMR vaccine: Using psychological risk factors to inform targeted communication strategies. Political Research Quarterly, 74(2), 464-478.

Mayweg-Paus, E., & Jucks, R. (2015). Evident or doubtful? How lexical hints in written information influence laypersons' understanding of influenza. Psychology, health & medicine, 20(8), 989-996.

Meppelink, C. S., van Weert, J. C., Haven, C. J., & Smit, E. G. (2015). The effectiveness of health animations in audiences with different health literacy levels: an experimental study. Journal of Medical Internet Research, 17(1), e3979.

Moran, M. B., Lucas, M., Everhart, K., Morgan, A., & Prickett, E. (2016). What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment. Journal of Communication in Healthcare, 9(3), 151-163.

Oh, S., Yi, Y. J., & Worrall, A. (2012). Quality of health answers in social Q&A. Proceedings of the American Society for Information Science and Technology.

Oh, S., Zhang, Y., & Park, M. S. (2016). Cancer information seeking in social question & answer services: Identifying health-related topics in cancer questions on Yahoo! Answers. Information Research, 21(3), n3.

Pelaccia, T., Tardif, J., Triby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. Medical education online, 16(1), 5890.

Pian, W., Song, S., & Zhang, Y. (2020). Consumer health information needs: A systematic review of measures. Information Processing & Management, 57(2), 102077.

Reyna, V. F. (2012). Risk perception and communication in vaccination decisions: A fuzzy-trace theory approach. Vaccine, 30(25), 3790-3797.

Sbaffi, L., & Rowley, J. (2017). Trust and credibility in web-based health information: a review and agenda for future research. Journal of Medical Internet Research, 19(6), e218.

Shah, C., & Pomerantz, J. (2010). Evaluating and predicting answer quality in community QA. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 411-418).

Sharon, A. J., Yom-Tov, E., & Baram-Tsabari, A. (2020). Vaccine information seeking on social Q&A services. Vaccine, 38(12), 2691-2699. https://doi.org/10.1016/j.vaccine.2020.02.010

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. Science communication, 39(5), 621-645.

Wang, L., Xian, Z., & Du, T. (2022). The public information needs of COVID-19 vaccine: A study based on online Q&A communities and portals in China. Frontiers in psychology, 13, 961181. https://doi.org/10.3389/fpsyg.2022.961181

World Health Organization. (2019). Ten threats to global health in 2019. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

Xu, Z., & Guo, H. (2018). Using text mining to compare online pro-and anti-vaccine headlines: Word usage, sentiments, and online popularity. Communication studies, 69(1), 103-122.

Zhang, Y. (2014). Beyond quality and accessibility: Source selection in consumer health information searching. Journal of the Association for Information Science and Technology, 65(5), 911-927.

Ziebland, S. (2012). Health and illness in a connected world: How might sharing experiences on the internet affect people's health? The Milbank Quarterly, 90(2), 219-249. https://doi.org/https://doi.org/10.1111/j.1468-0009.2012.00662.x

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iconf (2025)

Barriers and facilitators to documenting social needs in electronic health records: a pilot survey

Kuo-Ting Huang, Ravi Patel, Sharon Connor, Elaine, Kong, Xuan Nie, and Jocelyn Schmersal

DOI: https://doi.org/10.47989/ir30iConf47152

Abstract

Introduction. This pilot study investigates the impact of socioeconomic status, perceived barriers, sociopsychological factors, and health-related contexts such as health literacy and the doctor-patient relationship on patient comfort and willingness to engage with electronic health records (EHRs).

Method. Data from 302 participants were collected via an online Amazon Mechanical Turk survey, assessing socioeconomic backgrounds, healthcare experiences, and willingness to share social needs and have them documented in EHR systems.

Analysis. Descriptive statistics analysed socioeconomic, sociopsychological, health-related, and outcome variables, followed by ordinary least squares regression to predict patients' willingness to share and document social needs in EHRs.

Results. The findings revealed that while a positive doctor-patient relationship increased patients' willingness to share social needs, it negatively influenced their willingness to have these needs documented in EHRs. In contrast, health literacy emerged as a significant positive predictor for willingness to document social needs in EHRs.

Conclusion. This study shows how the doctor-patient relationship and health literacy impact patient engagement with EHRs. While patients are willing to share social needs, concerns about data privacy hinder documentation. Improving health literacy and addressing socioeconomic disparities are crucial for enhancing documentation practices, promoting more patient-centred and equitable healthcare through digital platforms.

Introduction

As digital health continues to transform the healthcare landscape, the integration of electronic health records (EHR) and associated platforms becomes a crucial tool for advancing health equity. This integration facilitates a deeper understanding of the diverse patient populations' willingness to share and document social determinants of health (SDOH), including economic stability, education access and quality, health care access and quality, neighborhood and built environment, social and community context (Office of Disease Prevention and Health Promotion). Initiatives like the Centers for Medicare and Medicaid Services' contemplation of quality measures for social risk screening highlight the potential for standardized EHR reporting to improve patient outcomes across various demographics (Albert et al., 2022; Billioux et al., 2017).

While SDOH are often considered community-wide issues, research also underscores their prevalence at the individual level, manifesting as unstable housing, food insecurity, transportation difficulties, and the need for utility payment assistance (Gruß et al., 2021; Kreuter et al., 2021; Tuzzio et al., 2022). Despite general openness to disclosing these personal challenges (Albert et al., 2022; Rogers et al., 2020), there remains significant hesitation to document these needs in EHR systems (Albert et al., 2022; De Marchis et al., 2019), especially among marginalized groups due to various perceived barriers (Drake et al., 2021). This study investigates factors that potentially affect patients' willingness to document SDOH in health records and prioritizes individual-level SDOH. By focusing on these aspects, the study seeks to advance healthcare practices that reduce health disparities and promote health equity.

Socioeconomic drivers in EHR participation

The integration of SDOH into patient care is increasingly recognized as essential for promoting health equity, particularly through digital health technologies like EHRs (Alley et al., 2016). Socioeconomic factors, such as income and education, significantly influence patient engagement with EHRs, impacting both comfort and usage (Adler & Stead, 2015; Davis et al., 2023). Therefore, understanding how socioeconomic factors, along with social needs, influence patients' willingness to share and have their needs documented in EHRs is crucial for addressing health disparities within digital healthcare frameworks.

Perceived barriers for patients to express social needs

Despite the technological advancements like EHRs facilitating SDOH screening, patients remain hesitant to share detailed social backgrounds information (Davidson et al., 2020; Garg et al., 2016). This reluctance can stem from concerns about data misuse (Herrera et al., 2019) or fears of stigmatization and clinician concern of causing embarrassment (Kepper et al., 2023). While it's evident that understanding patient perspectives is pivotal for effective healthcare, a gap remains. Traditional healthcare often overlooks the patient's viewpoint, potentially leading to ineffectual treatments (Singh Ospina et al., 2019). Despite general patient support for SDOH screenings, patients' willingness to disclose personal information varies across demographics, revealing barriers to full participation (Rogers et al., 2020).

Theory of planned behavior as a framework

To further understand the dynamics of patient engagement with EHRs, the theory of planned behaviour (TPB) provides a valuable framework. It explains how attitudes, subjective norms, and perceived behavioral control shape health-related behaviours, such as the willingness to share and document social needs (Ajzen, 1991). These three constructs are critical in influencing a patient's decision to disclose social determinants, guided by anticipated outcomes, societal expectations, and confidence in communication, as outlined by Hardeman et al. (2002). This framework has been applied in health behavior studies by Davis et al. (2023) and Zellmer et al. (2022). Applying TPB to social needs assessments helps explain patients' decisions to disclose or withhold information, based on their belief that sharing benefits health and aligns with societal norms (Ajzen, 1991).

Health care factors

Health literacy and the doctor-patient relationship are key elements in enhancing health equity through digital health communication. Health literacy extends beyond understanding medical language to encompass the ability and motivation to effectively use health information, crucial for understanding and expressing SDOH (Berkman et al., 2011; Nutbeam & Kickbusch, 1998). In digital health settings, health literacy significantly affects patients' readiness to share their social needs in EHRs (Nutbeam, 2008). Studies found that patients with limited health literacy face greater challenges in healthcare setting (Logan, 2017; Ungar et al., 1998). Considering their significant impact on patient engagement, this study also examines how health literacy and the doctor-patient relationship potentially influence patients' willingness to discuss and document their social needs in EHRs.

Research questions

Given the above contexts, the study addresses the following questions:

RQ1: how do socioeconomic factors such as income and education level influence patients' comfort and willingness to document social determinants of health in EHRs?

RQ2: how do barriers related to providers, the consultation environment, and perceived legitimacy impact the comfort level of patients in discussing social needs with clinicians and their subsequent willingness to have these needs documented in EHRs?

RQ3: in the context of the TPB, how do subjective norms, attitudes towards expressing social needs, and perceived behavioral control influence patients' comfort in sharing social needs and their inclination to have these needs documented in EHRs?

RQ4: how do health literacy and the doctor-patient relationship impact patients' willingness to discuss and document their social needs in EHRs?

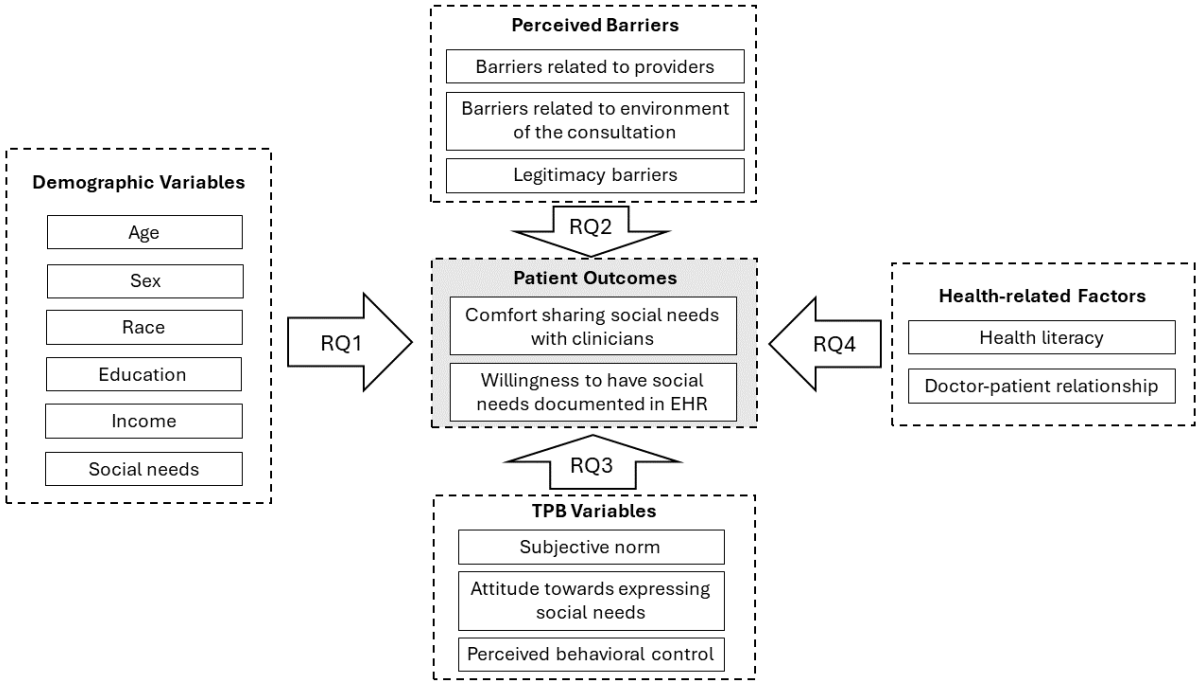


Figure 1. Conceptual model and research questions

Methods

Samples and design

To answer our research questions, 302 participants were sampled from Amazon Mechanical Turk (MTurk) in Spring 2023. MTurk is a popular online platform where people can perform tasks for a small financial reward (Paolacci & Chandler, 2014). It has been a valuable resource for research over the years (Goodman et al., 2013), and the data obtained through MTurk is generally regarded to be as good or better than data from undergraduate research groups (Miller et al., 2017). After excluding 20 participants for incomplete data, the final sample size was 282. Each participant received a $1 compensation.

Survey instruments and analytic tools

The survey assessed socioeconomic variables such as race, sex, income, and education levels, which are recognized as significant predictors of how patients interact with EHR systems. Besides the demographic variables, we adopted a 10-item measure from a previous study on patients' willingness to report social needs (Albert et al., 2022), based on the Accountable Health Communities Health-Related Social Needs Screening Tool (Billioux et al., 2017), covering housing instability, food insecurity, transportation, and utility assistance.

Subsequently, the survey examined perceived barriers to documenting SDOH in EHRs, employing a 16-item scale from Brandes et al. (2015) to identify issues related to healthcare providers' behavior, consultation environment, and the perceived legitimacy of EHR systems. The next section of the survey incorporated scales based on the Theory of Planned Behavior from Lin et al. (2016) to measure attitudes, subjective norms, and perceived behavioral control over sharing and documenting social needs. For healthcare factors, health literacy and the doctor-patient relationship were evaluated using validated scales from Chinn & McCarthy (2013) and Graffigna et al. (2017), providing insights into how patients understand and use health information and their interactions with providers.

Two outcomes were measured to evaluate the effectiveness of EHR integration in addressing social needs: (1) patients' comfort level in reporting social needs to clinicians, and (2) patients' willingness to document these needs in EHRs, using established frameworks from the studies by Schickedanz et al. (2019) and Albert et al. (2022). Descriptive statistics and Ordinary Least Squares (OLS) regressions, performed with SPSS version 29.0, analysed the impact of socioeconomic, sociopsychological, and health-related factors on these outcomes, offering detailed insights into the research questions.

Results

Descriptive analysis

Out of 302 survey participants, the average age was 36.48 years with a standard deviation of 10.75. Most participants (73.18%) were male, identified as White (94%), and held at least a college degree (88%). Income distribution among respondents showed 32.3% earning between $25,000 and $49,999, 35.1% between $50,000 and $74,999, and 24.5% earning between $75,000 and $99,999. Social needs assessment revealed that 60% of participants reported housing issues, 55% had concerns regarding utilities, 50.4% faced food insecurity, and 17.4% experienced transportation difficulties, with the range of reported social needs varying from 0 to 4.

In their interactions with healthcare providers, patients indicated moderate concerns across various perceived barriers, particularly regarding the appropriateness of discussing social needs. Analysis of the theory of planned behavior variables revealed that attitudes, behavioral control, and subjective norms moderately influenced the expression of social needs. Health literacy was assessed as relatively high, suggesting a good understanding and use of health information. The doctor-patient relationship was generally viewed positively, enhancing communication. However,

patients showed some hesitancy when it came to sharing social needs with clinicians, and even more reluctance to have these needs documented in EHRs, indicating potential concerns about privacy or the management of their information.

Patients generally feel comfortable sharing their social needs with clinicians, indicating a high level of openness during personal interactions. However, their willingness to have these social needs documented in electronic health records (EHRs) is more moderate, suggesting some reservations about the privacy or use of this sensitive information in digital formats. Cronbach's alpha (1951) tests confirmed the reliability of these measures, with all constructs' alpha values exceeding 0.7, indicating good internal consistency (Hair et al., 2010; Sürücü & Maslakci, 2020). Table 1 displays the descriptive statistics for the variables of interest.

Construct (Cronbach's α)	Mean (S.D.)	Min/Max
Barriers related to providers (0.851)	5.04 (1.17)	1/7
Barriers related to environment of the consultation (0.891)	4.96 (1.24)	1/7
Legitimacy barriers (0.868)	4.91 (1.47)	1/7
Subjective norm (0.838)	5.03 (1.36)	1/7
Attitude towards expressing social needs (0.897)	5.47 (1.07)	1/7
Perceived behavior control (0.837)	5.27 (1.16)	1/7
Health Literacy (N/A)	2.30 (0.37)	1/3
Doctor-Patient Relationship (0.926)	5.39 (0.85)	1/7
Comfort Sharing social needs with Clinicians (0.846)	3.71 (0.72)	0/4
Willingness to have social needs documented in EHR (0.767)	2.32 (1.51)	0/4
N=302		

Note: For 7-point scales, 1 is the lowest and 7 is the highest. For 3-point scales, 1 is the lowest and 3 is the highest. For 4-point scales, 0 is the lowest and 4 is the highest.

Table 1. Descriptive analysis of predictors and outcome variables

Ordinary least square (OLS) regression

The OLS regression models offer revealing insights into patients' comfort with sharing social needs (See Table 2). Regarding demographic variables, income was a significant predictor for both comfort in sharing social needs ($\beta = .070$, $p < .05$) and willingness to have them documented in EHR ($\beta = .181$, $p < .05$). This indicates that patients with higher income levels may be more comfortable and willing to engage in health processes involving their social needs. Financial stability could lead to greater confidence in interactions with healthcare systems. Social needs had a significant positive relationship with the willingness to have these needs documented in EHR ($\beta = .142$, $p < .05$). Patients acknowledging their social needs are likely more open to having them formally recognized and addressed within their health records.

Regarding the TPB variables, the attitude towards expressing social needs stood out as a significant factor for both comfort in sharing ($\beta = .343$, $p < .001$) and willingness to document ($\beta = .251$, $p < .05$) social needs in EHRs. This underscores the influence of individual beliefs on engagement with healthcare providers. Conversely, perceived behavioral control and subjective norms did not emerge as significant predictors, indicating that patients' confidence in their ability to express social needs and their perceptions of societal expectations may not be as impactful in this context. Similarly, the various barriers did not significantly predict participants' comfort or willingness to document social needs in EHRs.

The doctor-patient relationship had a nuanced impact, enhancing comfort with sharing social needs ($\beta = .149$, $p < .01$) while correlating with a lower willingness to document these needs in EHRs ($\beta = -.370$, $p < .05$). This suggests that when patients have a trusting relationship with their doctors,

they may feel less need for formal documentation of their social needs, relying instead on the strength of their interpersonal communication. Health literacy's significant positive correlation with the willingness to document social needs ($\beta = .957$, $p < .001$) indicates that patients who better comprehend health information may advocate more for integrating their social needs into medical records, reflecting a desire for a more holistic approach to health care management. These findings highlight the roles of income, attitudes, health literacy, and trust in shaping patients' comfort and willingness to document social needs, pointing to the need for education and communication strategies.

	Comfort Sharing Social Needs with Clinicians				**Willingness to have social needs documented in EHR**			
Demographic Variables	**Coefficient (SE)**		**95% CI**		**Coefficient (SE)**		**95% CI**	
Age	.002	(.003)	-.003	.007	.004	(.008)	-.013	.020
Sex (female as reference)	-.066	(.059)	-.183	.050	.050	(.192)	-.327	.427
Race (non-white as reference)	-.068	(.113)	-.289	.154	.098	(.364)	-.619	.815
Education	-.025	(.030)	-.084	.034	.076	(.097)	-.115	.267
Income	**.070***	(.027)	.017	.123	**.181***	(.087)	.009	.353
Social needs	.035	(.020)	-.004	.074	**.142***	(.065)	.015	.269
Perceived Barriers	**Coefficient (SE)**		**95% CI**		**Coefficient (SE)**		**95% CI**	
Barriers related to providers	.099	(.051)	-.002	.200	-.081	(.166)	-.408	.245
Barriers related to environment	-.015	(.049)	-.111	.082	.203	(.158)	-.108	.515
Legitimacy barriers	.033	(.029)	-.024	.090	.131	(.094)	-.053	.316
TPB variables	**Coefficient (SE)**		**95% CI**		**Coefficient (SE)**		**95% CI**	
Subjective norm	-.027	(.033)	-.092	.037	-.170	(.107)	-.380	.040
Attitude towards expressing	**.343*****	(.039)	.267	.420	**.251***	(.126)	.003	.500
Perceived behavioral control	.029	(.047)	-.063	.120	.109	(.151)	-.188	.406
Health-related Factors	**Coefficient (SE)**		**95% CI**		**Coefficient (SE)**		**95% CI**	
Doctor-Patient Relationship	**.149****	(.046)	.059	.239	**-.370***	(.147)	-.661	-.080
Health Literacy	-.079	(.078)	-.233	.075	**.957*****	(.253)	.459	1.454
	F (14, 286) = 33.80***				F (14, 286) = 4.30***			
	Adjusted R^2 = .605				Adjusted R^2 = .133			
Note: N=302. *P<.05, **P<.01, ***P<.001								

Table 2. Predictors of patient comfort and willingness to document social needs in EHRs

Discussion

Summary and interpretation of the results

Our study highlights critical aspects influencing patient engagement with EHRs, especially in relation to social needs. The significant role of income, as identified in our study, echoes findings from previous studies (Albert et al., 2022; Davis et al., 2023), reflecting the socioeconomic barriers to digital health access and utilization, as discussed in earlier research (Adler & Stead, 2015). This study builds on the findings described by Albert, et.al. (2022) in that it goes beyond a single health care setting and system. This accentuates the potential barriers marginalized groups may encounter in digital healthcare environments.

We also identified that patients with higher health literacy are more likely to permit the documentation of their social needs. In a scoping review conducted by Wark et al. (2021) examined stakeholder (e.g., patient and community) engagement in integrating SDOH data into EHR systems. They noted that subjects were less likely to find the practice acceptable when communication was inadequate or when the subjects believed the social needs were not actionable by the provider, particularly among those with low health literacy. The review concluded that more information is needed to improve patients' engagement and determine acceptability. Further, our study results also align with previous findings (Gruß et al., 2021; Kreuter et al., 2021), underscoring the importance of integrating SDOH into digital health platforms for health equity. The study also revealed that personal attitudes toward expressing social needs, consistent with the TPB (Ajzen, 1991; Hardeman et al., 2002), indicating that these individual beliefs significantly influence patient engagement with healthcare providers.

Interestingly, our study reveals that the doctor-patient relationship plays a nuanced role in how patients engage with EHRs. While there is an increased willingness among patients to share their social needs with clinicians, there is a noticeable decline in their comfort with having these needs formally recorded in EHRs. This suggests that while trust in the doctor-patient relationship may encourage verbal disclosure, concerns about privacy and the usage of personal information may inhibit the formal documentation of these discussions. Additionally, patients with higher health literacy are more willing to have their social needs documented, underscoring the importance of health literacy in the process. This finding reflects insights from previous studies (Grünloh et al., 2018; Hall et al., 2001), suggesting that strong interpersonal communication may reduce the need for formal documentation. Furthermore, the correlation between high health literacy and documentation willingness (Berkman et al., 2011; Nutbeam & Kickbusch, 1998), highlights the pivotal role of health literacy in enabling patients to engage more effectively with EHRs, thus enhancing personalized and equitable care through EHRs.

Significance

This study offers significant theoretical contributions to the understanding of patient engagement with EHRs, particularly in sharing and documenting social needs information. By delineating '*discomfort with social needs information in the medical record*' into '*comfort with sharing such information with clinicians*' and '*willingness to have it documented*', it provides novel insights into patients' perspectives. This distinction not only broadens the theoretical landscape as discussed in the works of authors like Adler and Stead (2015) and Gruß et al. (2021), but also deepens our comprehension of the multifaceted factors that shape patients' openness to share sensitive health data. This pilot study informs future research on addressing health literacy gaps and the doctor-patient relationship, laying the foundation for broader studies and practical improvements in digital health equity.

The practical implications of these findings are crucial for improving healthcare practices. The doctor-patient relationship's impact on EHR usage underscores the importance of integrating structural competency and cultural humility into clinical practice (Lekas et al., 2020; Metzl & Hansen, 2014). Patient-centred communication requires cultural and structural sensitivity and is essential for achieving better health outcomes (Naughton, 2018; Rathert et al., 2013). EHR technologies should be culturally tailored to address gaps in SDOH screening, creating environments where patients feel comfortable and empowered to share their needs (Rathert et al., 2017). Implications for clinician training and patient education are vital in enhancing patient engagement with EHRs, as indicated by previous studies (Bhattad & Pacifico, 2022; Hartzler et al., 2018; Rogers et al., 2020). Developing these areas could significantly improve patients' interaction with EHRs.

Additionally, by examining barriers to sharing social needs and applying the TPB, our study sheds light on potential interventions that could promote open communication between patients and

healthcare providers. Enhancing provider awareness of patient needs and perceptions emphasizes the importance of aligning documentation with patient-centred care in EHRs (Stanhope & Matthews, 2019). These interventions are vital for addressing the barriers encountered by marginalized groups, aligning with the overarching aim of advancing health equity in digital health technologies. The insights from this study can assist healthcare providers and policymakers in developing strategies that ensure equitable access to digital health resources, contributing significantly to reduce health disparities.

Limitation and future research

This study's limitations include its reliance on an MTurk sample, suggesting a higher socioeconomic status due to the requirement for internet access. In addition, despite investigating social needs at the individual levels, we could not fully explore their diverse effects across different populations, potentially impacting data interpretation. Based on this pilot study, future research should include a representative sample to better understand the factors influencing patients' willingness to share and document social needs in EHRs across different demographics and healthcare settings. Furthermore, the lack of data on participants' medical conditions limits insight into how these might affect their willingness to share social needs. Additionally, this study highlights the critical roles of doctor-patient relationships, health literacy, and socioeconomic factors. However, our sample has limited power to fully explore the needs across various populations. Future research should develop interventions tailored to diverse underserved communities, aimed at enhancing these interactions, improving health literacy, and addressing socioeconomic disparities across varied healthcare environments and cultural contexts.

Conclusion

This research highlights the critical roles of the doctor-patient relationship and health literacy in determining how patients engage with EHRs. The findings demonstrate that a positive interaction with healthcare providers encourages patients to share their social needs, although there is reluctance to have these needs formally documented due to concerns about privacy and personal data usage. This finding also shows the importance of enhancing health literacy, which correlates with a greater willingness to document social needs in EHRs. Additionally, addressing socioeconomic disparities that affect these interactions is crucial for improving patient-centred care. Collectively, these efforts are essential for enriching healthcare experiences and promoting more equitable and effective healthcare delivery through digital platforms.

Acknowledgements

The authors acknowledge support from the School of Computing and Information at the University of Pittsburgh through the start-up fund provided to the corresponding author.

About the authors

Kuo-Ting (Tim) Huang is an Assistant Professor in the School of Computing and Information, University of Pittsburgh, Pennsylvania, USA. He received his Ph.D. from Michigan State University. His research investigates the potential of emerging technologies to bridge gaps in health and education for underserved and disadvantaged populations. He can be contacted at timhuang@pitt.edu

Ravi Patel, PharmD, MBA, MS, is the Lead Innovation Advisor at the University of Pittsburgh School of Pharmacy, where he leads the Pharmacy Innovation Lab. He received his Doctor of Pharmacy (PharmD) from the University of Pittsburgh. His research interests include user-centered design of health technology, evaluation of health data, and the application of creativity and innovation in healthcare education and practice. He can be contacted at rmp40@pitt.edu

Sharon Connor is an Associate Professor of Pharmacy and Therapeutics at the School of Pharmacy, University of Pittsburgh. She received her PharmD from Creighton University in Omaha, Nebraska. She collaborates with community partners to build pharmacy programs designed to enhance access to essential medications. Her community engaged research focuses on innovative programs that promote health in resource-limited settings. She can be reached at sconnor@pitt.edu

Elaine Kong is a PhD student in the School of Computing and Information at the University of Pittsburgh, Pennsylvania, USA. Her research interests focus on how cancer survivors in underserved communities seek information and the challenges they face in finding reliable, evidence-based resources. She can be contacted at elainekong@pitt.edu

Xuan Nie is a PhD student in the School of Computing and Information at the University of Pittsburgh, Pennsylvania, USA. Her research focuses on helping seniors improve their self-management of health conditions, including health information access, information behavior research, and interventions for improvement. She can be contacted at xun9@pitt.edu

References

Adler, N. E., & Stead, W. W. (2015). Patients in context—EHR capture of social and behavioral determinants of health. Obstetrical & Gynecological Survey, 70(6), 388-390.

Ajzen, I. (1991). The theory of planned behavior. Organizational behavior and human decision processes, 50(2), 179-211.

Albert, S. M., McCracken, P., Bui, T., Hanmer, J., Fischer, G. S., Hariharan, J., & James III, A. E. (2022). Do patients want clinicians to ask about social needs and include this information in their medical record? BMC health services research, 22(1), 1275.

Alley, D. E., Asomugha, C. N., Conway, P. H., & Sanghavi, D. M. (2016). Accountable health communities—addressing social needs through Medicare and Medicaid. N Engl J Med, 374(1), 8-11.

Beach, M. C., Price, E. G., Gary, T. L., Robinson, K. A., Gozu, A., Palacio, A., Smarth, C., Jenckes, M. W., Feuerstein, C., Bass, E. B., Powe, N. R., & Cooper, L. A. (2005). Cultural Competence: A Systematic Review of Health Care Provider Educational Interventions. Medical care, 43(4), 356-373. https://doi.org/10.1097/01.mlr.0000156861.58905.96

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., & Crotty, K. (2011). Low Health Literacy and Health Outcomes: An Updated Systematic Review. Annals of internal medicine, 155(2), 97-107. https://doi.org/10.7326/0003-4819-155-2-201107190-00005

Bhattad, P. B., & Pacifico, L. (2022). Empowering patients: promoting patient education and health literacy. Cureus, 14(7).

Billioux, A., Verlander, K., Anthony, S., & Alley, D. (2017). Standardized screening for health-related social needs in clinical settings: the accountable health communities screening tool. NAM perspectives.

Brandes, K., Linn, A. J., Smit, E. G., & van Weert, J. C. (2015). Patients' reports of barriers to expressing concerns during cancer consultations. Patient Education and Counseling, 98(3), 317-322.

Chinn, D., & McCarthy, C. (2013). All Aspects of Health Literacy Scale (AAHLS): developing a tool to measure functional, communicative and critical health literacy in primary healthcare settings. Patient Education and Counseling, 90(2), 247-253.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. psychometrika, 16(3), 297-334.

Das, D., & Sengar, A. (2022). A fuzzy analytic hierarchy process-based analysis for prioritization of barriers to the adoption of eHealth in India. International Journal of Medical Informatics, 165, 104830.

Davidson, K. W., Kemper, A. R., Doubeni, C. A., Tseng, C.-W., Simon, M. A., Kubik, M., Curry, S. J., Mills, J., Krist, A., & Ngo-Metzger, Q. (2020). Developing primary care–based recommendations for social determinants of health: methods of the US Preventive Services Task Force. Annals of internal medicine, 173(6), 461-467.

Davis, V. H., Dainty, K. N., Dhalla, I. A., Sheehan, K. A., Wong, B. M., & Pinto, A. D. (2023). "Addressing the bigger picture": A qualitative study of internal medicine patients' perspectives on social needs data collection and use. Plos one, 18(6), e0285795.

De Marchis, E. H., Hessler, D., Fichtenberg, C., Adler, N., Byhoff, E., Cohen, A. J., Doran, K. M., de Cuba, S. E., Fleegler, E. W., & Lewis, C. C. (2019). Part I: a quantitative study of social risk screening acceptability in patients and caregivers. American journal of preventive medicine, 57(6), S25-S37.

Drake, C., Batchelder, H., Lian, T., Cannady, M., Weinberger, M., Eisenson, H., Esmaili, E., Lewinski, A., Zullig, L. L., & Haley, A. (2021). Implementation of social needs screening in primary care: a qualitative study using the health equity implementation framework. BMC health services research, 21, 1-16.

Garg, A., Boynton-Jarrett, R., & Dworkin, P. H. (2016). Avoiding the unintended consequences of screening for social determinants of health. Jama, 316(8), 813-814.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. Journal of Behavioral Decision Making, 26(3), 213-224.

Graffigna, G., Barello, S., Bonanomi, A., & Riva, G. (2017). Factors affecting patients' online health information-seeking behaviours: The role of the Patient Health Engagement (PHE) Model. Patient Education and Counseling, 100(10), 1918-1927.

Grünloh, C., Myreteg, G., Cajander, Å., & Rexhepi, H. (2018). "Why do they need to check me?" Patient participation through eHealth and the doctor-patient relationship: qualitative study. Journal of medical Internet research, 20(1), e11.

Gruß, I., Varga, A., Brooks, N., Gold, R., & Banegas, M. P. (2021). Patient interest in receiving assistance with self-reported social risks. The Journal of the American Board of Family Medicine, 34(5), 914-924.

Hair, J., Anderson, R., Babin, B., & Black, W. (2010). Multivariate data analysis: A global perspective (7): Pearson Upper Saddle River. In: NJ.

Hall, M. A., Dugan, E., Zheng, B., & Mishra, A. K. (2001). Trust in Physicians and Medical Institutions: What Is It, Can It Be Measured, and Does It Matter? The Milbank quarterly, 79(4), 613-639. https://doi.org/10.1111/1468-0009.00223

Hardeman, W., Johnston, M., Johnston, D., Bonetti, D., Wareham, N., & Kinmonth, A. L. (2002). Application of the theory of planned behaviour in behaviour change interventions: A systematic review. Psychology and health, 17(2), 123-158.

Hartzler, A. L., Tuzzio, L., Hsu, C., & Wagner, E. H. (2018). Roles and functions of community health workers in primary care. The Annals of Family Medicine, 16(3), 240-245.

Herrera, C.-N., Brochier, A., Pellicer, M., Garg, A., & Drainoni, M.-L. (2019). Implementing social determinants of health screening at community health centers: clinician and staff perspectives. Journal of primary care & community health, 10, 2150132719887260.

Kepper, M. M., Walsh-Bailey, C., Prusaczyk, B., Zhao, M., Herrick, C., & Foraker, R. (2023). The adoption of social determinants of health documentation in clinical settings. Health services research, 58(1), 67-77.

Kreuter, M. W., Thompson, T., McQueen, A., & Garg, R. (2021). Addressing social needs in health care settings: evidence, challenges, and opportunities for public health. Annual review of public health, 42(2021), 329-344.

Lekas, H.-M., Pahl, K., & Fuller Lewis, C. (2020). Rethinking cultural competence: Shifting to cultural humility. Health services insights, 13, 1178632920970580.

Lin, C.-Y., Updegraff, J. A., & Pakpour, A. H. (2016). The relationship between the theory of planned behavior and medication adherence in patients with epilepsy. Epilepsy & Behavior, 61, 231-236.

Logan, R. A. (2017). Seeking an expanded, multidimensional conceptual approach to health literacy and health disparities research. Information Services & Use, 37(1), 59-83.

Matusitz, J., & Spear, J. (2014). Effective doctor–patient communication: an updated examination. Social work in public health, 29(3), 252-266.

Metzl, J. M., & Hansen, H. (2014). Structural competency: theorizing a new medical engagement with stigma and inequality. Social science & medicine, 103, 126-133.

Miller, J. D., Crowe, M., Weiss, B., Maples-Keller, J. L., & Lynam, D. R. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. Personality Disorders: Theory, Research, and Treatment, 8(1), 26.

Naughton, C. A. (2018). Patient-centered communication. Pharmacy, 6(1), 18.

Nutbeam, D. (2008). The evolving concept of health literacy. Social science & medicine (1982), 67(12), 2072-2078. https://doi.org/10.1016/j.socscimed.2008.09.050

Nutbeam, D., & Kickbusch, I. (1998). Health promotion glossary. Health promotion international, 13(4), 349-364.

Office of Disease Prevention and Health Promotion. (n.d.). Social determinants of health. Healthy People 2030. U.S. Department of Health and Human Services. Retrieved September 4, 2024. from https://health.gov/healthypeople/priority-areas/social-determinants-health

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. Current directions in psychological science, 23(3), 184-188.

Rathert, C., Mittler, J. N., Banerjee, S., & McDaniel, J. (2017). Patient-centered communication in the era of electronic health records: What does the evidence say? Patient Education and Counseling, 100(1), 50-64.

Rathert, C., Wyrwich, M. D., & Boren, S. A. (2013). Patient-centered care and outcomes: a systematic review of the literature. Medical Care Research and Review, 70(4), 351-379.

Rogers, A. J., Hamity, C., Sharp, A. L., Jackson, A. H., & Schickedanz, A. B. (2020). Patients' attitudes and perceptions regarding social needs screening and navigation: multi-site survey in a large integrated health system. Journal of general internal medicine, 35, 1389-1395.

Schickedanz, A., Hamity, C., Rogers, A., Sharp, A. L., & Jackson, A. (2019). Clinician experiences and attitudes regarding screening for social determinants of health in a large integrated health system. Medical care, 57(Suppl 6 2), S197.

Singh Ospina, N., Phillips, K. A., Rodriguez-Gutierrez, R., Castaneda-Guarderas, A., Gionfriddo, M. R., Branda, M. E., & Montori, V. M. (2019). Eliciting the patient's agenda-secondary analysis of recorded clinical encounters. Journal of general internal medicine, 34, 36-40.

Stanhope, V., & Matthews, E. B. (2019). Delivering person-centered care with an electronic health record. BMC medical informatics and decision making, 19, 1-9.

Sürücü, L., & Maslakci, A. (2020). Validity and reliability in quantitative research. Business & Management Studies: An International Journal, 8(3), 2694-2726.

Tuzzio, L., Wellman, R. D., De Marchis, E. H., Gottlieb, L. M., Walsh-Bailey, C., Jones, S. M., Nau, C. L., Steiner, J. F., Banegas, M. P., & Sharp, A. L. (2022). Social risk factors and desire for assistance among patients receiving subsidized health care insurance in a US-based integrated delivery system. The Annals of Family Medicine, 20(2), 137-144.

Ungar, W. J., Coyte, P. C., Borden, E. K., Gaebel, K. A., Goldsmith, C. H., Hu, X., Levine, M. A., Poston, J. W., Willison, D. J., & Board, P. M. M. P. A. (1998). Health services utilization reporting in respiratory patients. Journal of clinical epidemiology, 51(12), 1335-1342.

Wark, K., Cheung, K., Wolter, E., & Avey, J. P. (2021). Engaging stakeholders in integrating social determinants of health into electronic health records: a scoping review. International journal of circumpolar health, 80(1), 1943983.

Zellmer, L., Johnson, B., Idris, A., Mehus, C. J., & Borowsky, I. W. (2022). Post-Identification approaches to addressing health-related social needs in primary care: a qualitative study. Journal of general internal medicine, 37(4), 802-808.

Zhang, H. (2022). The shift in the narrative of doctor-patient communication and the cultivation of medical information exchange communication based on the information technology era. Mobile Information Systems, 2022, 1-11.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

The challenges of cross-organizational patient data sharing in medical consortium: a study within the context of two-way referral systems

Lihong Zhou, Hanyu Miao, Ruiyang Tu, and Jiangfeng Hu

DOI: https://doi.org/10.47989/ir30iConf47149

Abstract

Objective. Under the background of two-way referral in medical consortium, the provision of high-quality hierarchical care requires active integration and cross-organizational sharing of patient data. However, cross-organizational data sharing among hospitals is not always spontaneous, comprehensive, and effective.

Methods. This paper reports on a research study aiming to identify the barriers of patient data sharing. Due to the exploratory nature of this study, two medical consortiums in central China were chosen to conduct a case study. Specifically, a mixed qualitative-quantitative approach was employed for the case study analysis and verification.

Conclusions. The analysis of qualitative interview data pointed to 27 barriers of data sharing under three main themes: psychological ownership; territorial behaviours; cross-organizational dynamism. The quantitative Kano analysis revealed the manifestation of these barriers, and their prioritization were classified into four types: must-be, one-dimensional, attractive, and indifference.

Implications. This study provides valuable references for enhancing the effectiveness of patient data sharing and optimizing the allocation of healthcare resources. The case setting is based on China's medical consortium, but the findings offer useful insights and indications that can be shared across international borders.

Introduction

Data, commonly understood as statistical observations or collected evidence (Zins, 2007), has become a crucial asset in modern societies (Hilbert, 2016). However, data is often only used to a limited extent or frequently only in an internal organization context (Prajogo et al., 2018). Data sharing refers to breaking down the data barriers between organizational departments, establishing a unified data sharing mechanism, and accelerating the flow of data resources within the organization (Harris et al., 2007). It also encompasses making research data publicly available with minimal reuse restrictions, known as open data (Thoegersen & Borlund, 2022). While data sharing increases the complexity and quantity of data (Kowalczyk & Shankar, 2011), it plays a vital role in advancing scientific collaboration (Priego et al., 2022), urban governance (Palm et al., 2019), public health management (Brand et al., 2022), and social development.

The hospitals are typically data-intensive organizations. Health decision making has become increasingly complex, and the use of patient data has become essential in this information age (Van, 2014). Integration of patient data has been perceived to be a breakthrough for the discovery of timely, cost-efficiency therapies, and a crucial step towards enhancing the quality of medical care for curing diseases (Weitzman et al., 2010). Two-way referral serves as a classic example of patient data sharing. It refers to the transfer process between hospitals of different levels, specialized hospitals, or between general and specialized hospitals (Huang et al., 2023) based on the health needs of the population. Nowadays, the integration and cross-organizational sharing of patient data has become a key part of achieving the goals of hierarchical diagnosis and two-way referral.

Recognizing the importance of addressing these medical needs and challenges, the Chinese government attempted to improve medical development through the establishment of medical consortium (Li et al., 2017). Medical consortium includes urban medical groups led by tertiary public hospitals, county medical communities led by county hospitals, and specialist alliances and telemedicine collaboration networks with specialist associations as the link (Cai et al., 2018).

Nevertheless, barriers exist to integrated working, including technical and ethical issues, organizational resistance, financial costs, and privacy concerns (Houtkoop et al., 2018; Li & Cong, 2024; Reddy et al., 2024). Meanwhile, different hospitals have different regulations and willingness on sharing patient data in the two-way referral process, which has made the implementation of hierarchical diagnosis and treatment less effective than expected. Yet current research in this field is notably limited.

The paper reports on a research project that aimed to identify and solve existing barriers to cross-organizational sharing of patient data in the context of two-way referral in medical consortium. To further the analysis, the following research questions were formulated:

RQ1. What are the barriers of patient data sharing?

RQ2. Which are the priorities of these barriers?

RQ3. What are the practical strategies to address these barriers?

Research methodology and design

To answer our research questions, and considering the exploratory nature of this study, a case study approach is employed. According to China's current '*Hospital Classification and Management Measures*', 3A hospitals are the highest level in the classification of hospitals in Mainland China. After careful consideration, Xiangyang First People's Hospital Medical Consortium and Xiangyang Traditional Chinese Medicine Hospital Medical Consortium is adopted as the research case. Both hospitals are Grade 3A hospitals, which ensures the representativeness of the research and can accurately reflect the situation in the entire medical field.

The research was carried out in two stages. In *stage one*, qualitative research was undertaken to identify the barriers of patient data sharing. Semi-structured interviews were used to gather qualitative data, with guaranteed access to interview respondents. All interviews were recorded using a digital recorder and transcribed verbatim into MS Word files. NVivo 14 was used to facilitate the analysis of interview data.

After that, a Kano model analysis was carried out in *stage two* to verify and validate the barriers identified in the qualitative study. In this stage, Kano model analysis was used to prioritize barriers of patient data sharing. It was expected that through the prioritization processed, effective strategic suggestions could be formulated.

Stage one: qualitative research

Research methods and processes

As discussed previously, semi-structured interviews were used to collect qualitative data, which were then analysed using a thematic analysis approach. A framework proposed by Brown et al. (2014) was used as a basis for data collection and analysis, as shown in Figure 1. The framework comprises three categories which are psychological ownership, territorial behaviours, and cross-organizational dynamism (Brown et al., 2014), that can help understanding barriers in patient data sharing.

- Psychological ownership: hospitals have invested a lot of manpower, material resources and time costs to maintain patient data, so that they have regarded patient data as their inseparable assets.
- Territorial behaviours: are actions can be adopted by hospitals for the protection of their perceived ownership over data. Hospitals are inclined to keep and safeguard their data instead of sharing it willingly.
- Cross-organizational dynamism: can moderate the influence of psychological ownership on territorial behaviours.

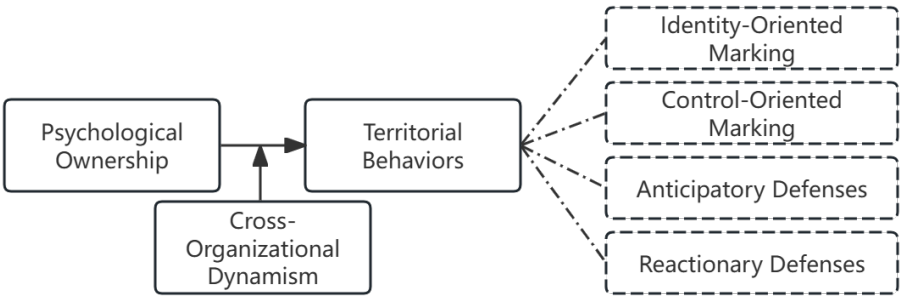


Figure 1. A preliminary analytical model based on the research of Brown et al. (2014)

The semi-structured question script was designed based on Figure 1 and consisted of 15 questions, including 1 ice-breaking question and 14 theoretical probing questions. Before entering the case study, the question script was proofread and trialled with two professors in the field of cross-organizational data sharing. A total of 13 staff and patients were approached and interviewed. The demographic profile of the participants is illustrated in the Table 1 below:

No.	Interviewer	Department/Section
A	Engineer	Information Center
B	Engineer	Information Center
C	Department Leader	Information Center
D	Clerk	Medical Office
E	Clerk	Medical Office
F	Department Leader	Information Center
G	Engineer	Information Center
H	Department Leader	Medical Office
I	Head Nurse	Nephrology
J	Nurse	Nephrology
K	Chief Physician	Nephrology
L	Referral patient	Orthopedics
M	Referral patient	Neurosurgery

Table 1. Information on interview participants

All interviews were interviewed in March 2024. At the beginning of every interview, the participant was consulted if the use of digital recorder could be accepted. The participants were assured that they can stop the recorder whenever they feel uncomfortable.

Instead of directly diving into the question-answer dialogue, the interviews started by making some irrelevant and light conversations. After obtaining informed consent, all interview recordings were then transcribed verbatim by the research team. The transcripts were analysed using a thematic analysis approach.

In the research, the Brown et al. (2014)'s framework was considered as generic requirements and were converted into a preliminary set of codes in the early stages. Throughout the data analysis, these codes were reconceptualized, expanded and revised, according to the statements, interpretations, and perspectives of the individual interview participants.

Findings

The analysis of the interview data inherited the thematic analysis strategy of the model (Figure 1), the themes and codes resulting from the analysis are shown in the Table 2.

Theme	Sub-theme	Coding	Identifier
Psychological Ownership	-	Fear of declining status after sharing	A1
		Reluctance to give up data assets	A2
		Lack of access to required data	A3
Territorial Behaviours	Identity-Oriented Marking	Using different data systems	B1
		Different data processing standards	B2
	Control-Oriented Marking	Complex approval process for declarations	C1
		Self-determination of the form of sharing	C2
	Anticipatory Defences	Distrust of shared patient data	D1
		Passive response to data sharing construction	D2
		Insufficient resources in primary medical institutions	D3
		Lack of personnel in data sharing positions	D4
	Reactionary Defences	Excessive accountability for data accidents	E1
Cross-Organizational Dynamism	Cross-Organizational Management	Deviation between government planning and real needs	F1
		Lack of clear policies and regulations	F2
		Lack of government supervision and guidance	F3
		Limited results of medical insurance reform	F4
	Cross-Organizational Relationships	Lack of high-level cross-organizational mechanism	G1
		Lack of preliminary cooperation foundation	G2
		Lack of trust between organizations	G3
		Interest game of tertiary hospitals	G4
		Loose cooperation of medical consortium	G5
	Cross-Organizational Risk	Unclear rights and responsibilities for sharing	H1
		Lack of incentive and compensation mechanism after sharing	H2
		Low level of informatization of medical institutions	H3
		Lack of financial support	H4
		Concerns about data quality	H5
		Concerns about data security	H6

Table 2. Requirements, themes, coding, and identifier

The data analysis shows that the interviewees frequently highlighted psychological barriers related to the perceived ownership of data assets. A significant challenge for hospitals is the fear of diminished status after data sharing. This concern is particularly pronounced among community and secondary hospitals, as an interviewee noted, '*liberalising data sharing will enable doctors at the next level of care to improve their standards, which poses a threat to us*' (Interview F). Besides, in the context of big data, '*data are considered assets*' (Interview A). Hospitals invest considerable resources, including manpower, material, and time, into maintaining patient data, treating it as a crucial and inseparable asset. As a result, they are unwilling to engage with or learn how to use data-sharing platforms, which further complicates their ability to access data through these platforms.

The data analysis showed that territorial behaviours can significantly affect patient data sharing, with four specific manifestations identified. Identity-oriented marking refers to the fact that '*each*

hospital has its own HIS system' (Interviewee B), enhancing the complication of integrating systems to achieve data sharing. Besides, '*data processing standards*' (Interviewee A) varies, with no unified stipulation to realize data cross-authentication. Control-oriented marking is evident in the lengthy approval processes of data sharing request, with '*several people need to sign only a piece of file and taking at least three working days to approve it*' (Interview M). Anticipatory defences especially reflect a passive response and prudent attitude to obtained data. Meanwhile, deliberate procrastination can also be a common action to defend settled data sharing request, since '*they even do not arrange people to specifically designated to handle this issue*' (Interview E). Reactionary defences generally occur when hospitals actually experience data infringement by others. There is evidence showing that an excessive accountability for data accidents will be taken as a defensive action to the specific hospital. As a doctor stated that '*if they fail to take the necessary security measures and cause a data breach, they must take the responsibility*' (Interview L).

Furthermore, the interviews revealed significant barriers to cross-organizational data sharing, primarily related to management, relationships, and risk. The most important barrier for management issue is the current legal framework lacks clarity, as noted by an interviewee, '*the overall plan does not conform to the actual situation*' (Interview G). To enable effective data sharing management, a unified data centre is essential. However, the absence of a high-level cross-organizational mechanism exacerbates this problem. In terms of relationships, hospitals express concerns about '*data security*' (Interview B/C/J) and '*data quality*' (Interview H/I/J), which affect the trust and collaboration for data sharing. Additionally, issues following data sharing include '*unclear sharing rights and responsibilities*' (Interview F) and an '*absence of an incentive compensation mechanism*' (Interview A). These risks highlight the need for clearer guidelines and effective incentives.

Stage two: Kano model analysis

Research methods and processes

Kano model is composed of two important concepts, one is the objective realization of requirements, the other is the subjective satisfaction of users. *Stage two* employs the Kano model to validate and verify the barriers identified in *stage one*. Moreover, *stage two* measures the priority of the barriers and classifies the barriers into four categories (sorted in the order they should be solved): Must-be (M); One-dimensional (O); Attractive (A); Indifferent (I) (Xu et al., 2009). Data collection is divided into three main steps.

- Questionnaire design. This study transforms the identified disorders into demand items. Excerpts from the interview transcripts from Stage One were used as evidence for the design of questionnaire.
- Grading of questions. Every option setting refers to the Likert 5-point scale method (Adelson & McCoach, 2010). And respondents can choose from one of the following five indicators: I like it that way, it must be that way, I am neutral, I can live with it that way, and I dislike it that way.
- Questionnaire distribution. This study uses both online and offline methods to place and collect questionnaires. The questionnaires were distributed and returned in April 2024, yielding 98 valid responses.

Findings

Defining primary quality attribute (PQA) was the first step in Kano model analysis. In the analysis, responses from the questionnaire survey to the functional and non-functional questions of the requirements were combined and then assigned to a Kano evaluation matrix (Table 3). Using the matrix, a requirement can be defined and categorized into one of the M, O, A, I, R and Q categories. For instance, the analysis of requirements B1 is shown in Table 4. B1 is an A requirement because A

represents the highest ratio. Repeating the above analysis processes, all requirements could be calculated.

Requirements		Non-functional issues				
	Scale	I dislike it that way	I can live with it that way	I am neutral	It must be that way	I like it that way
Functional questions	I dislike it that way	Q	R	R	R	R
	I can live with it that way	M	I	I	I	R
	I am neutral	M	I	I	I	R
	It must be that way	M	I	I	I	R
	I like it that way	O	A	A	A	Q

Table 3. Kano evaluation matrix

Requirements	A	O	M	I	R	Q
B1	48.0%	10.2%	18.4%	23.5%	0	0

Table 4. A sample of Kano evaluation matrix for B1

User satisfaction coefficient (UCS) measures a situation whereby achieving a requirement of patient data sharing can prevent dissatisfaction. It comprises two indicators: satisfaction index (SI) and dissatisfaction index (DSI).

$$SI = \frac{A + O}{A + O + M + I} \quad (1)$$

$$DSI = \frac{(-1) \times (O + M)}{A + O + M + I} \quad (2)$$

Based on the calculation of SI and DSI, a satisfaction impact matrix can be developed with an X-axis presenting the DSI values and a Y-axis illustrating the SI values.

$$X = DSI(DSI_{min}, DSI_{max}) DSI[-1,0] \quad (3)$$

$$Y = SI(SI_{min}, SI_{max}) SI[0,1] \quad (4)$$

$$O(X, Y) = (X_{ave}, Y_{ave}) \quad (5)$$

Therefore, the UCS is developed as shown in Figure 2. Moreover, the barriers are divided into four sections: Attractive (A): B1, C1, D4, E1, F2, G2, G4, G5, H4. One-dimensional (O): B2, H2, H6. Must-be (M): A2, A3, D2, F1, G1, H1, H3, H5. Indifferent (I): A1, C2, D1, D3, F3, F4, G3.

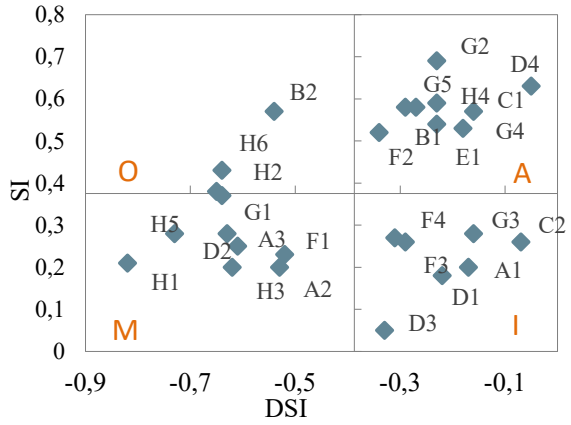


Figure 2. User satisfaction impact matrix

Discussions and conclusions

This paper uses a comprehensive approach to identify barriers to patient data sharing and validate the prioritization of barriers. Using deductive qualitative methods, 27 barriers of data sharing were identified under three main themes. Subsequently, these barriers were categorized into four levels using Kano analysis based on the quantitative measurement. This study generates the following specific strategic suggestions.

First, establishing standardized patient data specifications is a critical prerequisite for effective data sharing. It is recommended to systematically categorize data types, encompassing essential attributes such as format, storage methods, and access rights. A comprehensive and compatible standardization framework should be developed. Based on this, a specialized team should be established to oversee the implementation of these data standards. Additionally, hospitals should prioritize the development of new technologies and increase capital investment to ensure the real-time updating and seamless transmission of critical medical information.

Second, to enhance the diagnostic and therapeutic capabilities of medical consortium, it is crucial to prioritize comprehensive training programs to ensure all employees are equipped with the essential skills and knowledge. This training should focus on fostering a robust understanding of collaboration and resource sharing within the network. As this awareness deepens, employees will gain a deeper understanding of data sharing. This understanding will ultimately foster a cooperative culture, enabling smooth and effective sharing of medical data and resources across organizations.

Third, to address the challenges of financial support and benefit distribution in data sharing, it is imperative to enhance government involvement. Specifically, the government should urgently increase financial support and establish a dedicated subsidy system for patient data sharing. Additionally, a cross-organizational communication framework centered on data sharing should be developed, with a committee comprising hospital representatives overseeing this initiative.

Fourth, for the sake of improving the staff's ability to understand and utilize the medical consortium effectively, hospitals should implement targeted educational activities emphasizing the importance of data quality, including holding lectures, seminars, etc. At the same time, hospitals should collaborate with data experts and software developers to design comprehensive training courses for the medical information system, covering system functions, operating logic, application scenarios, etc.

In conclusion, cross-organizational patient data sharing in medical consortium offers a unique opportunity for hospitals to redefine, reposition and reinvent themselves. This research adds to

the general knowledge base by proposing a comprehensive set of barriers of patient data sharing and strategic suggestions, which can be adopted as a knowledge base and a starting point for future studies. Also, future work can aim at expanding the sample coverage and conducting research on different types of medical consortium.

About the authors

Lihong Zhou is a Professor in School of Information Management, Wuhan University, Hubei, China. He received his Ph.D. from the University of Sheffield and his research interests are in library management and data sharing. He can be contacted at L.zhou@whu.edu.cn

Hanyu Miao is currently pursuing a master's degree in Library Science at School of Information Management, Wuhan University, Wuhan, China. Her research interests are in library management and data sharing. She can be contacted at Miaohyeah@163.com

Ruiyang Tu received his master's degree from School of Information Management, Wuhan University, Wuhan, China. His research interest is in data sharing. He can be contacted at trysim@whu.edu.cn

Jiangfeng Hu is a Ph.D. student in School of Information Management, Wuhan University, Hubei, China. Her research interests are in library management and data sharing. She can be contacted at jiangfeng.hu@whu.edu.cn

References

Adelson, J. L., & McCoach, D. B. (2010). Measuring the Mathematical Attitudes of Elementary Students: The Effects of a 4-Point or 5-Point Likert-Type Scale. Educational and Psychological Measurement, 70(5), 796-807. https://doi.org/10.1177/0013164410366694

Brand, D., Singh, J. A., McKay, A. G. N., Cengiz, N., & Moodley, K. (2022). Data sharing governance in sub-Saharan Africa during public health emergencies: Gaps and guidance. South African Journal of Science, 118(11-12), Article #13892. https://doi.org/10.17159/sajs.2022/13892

Brown, G., Crossley, C., & Robinson, S. L. (2014). PSYCHOLOGICAL OWNERSHIP, TERRITORIAL BEHAVIOR, AND BEING PERCEIVED AS A TEAM CONTRIBUTOR: THE CRITICAL ROLE OF TRUST IN THE WORK ENVIRONMENT. Personnel Psychology, 67(2), 463-485. https://doi.org/10.1111/peps.12048

Cai, M., Liu, E., Tao, H. B., Qian, Z. M., Fu, Q., Lin, X. J., . . . Ni, Z. L. (2018). Does A Medical Consortium Influence Health Outcomes of Hospitalized Cancer Patients? An Integrated Care Model in Shanxi, China. International Journal of Integrated Care, 18(2), Article 7. https://doi.org/10.5334/ijic.3588

Harris, D., Khan, L., Paul, R., & Thuraisingham, B. (2007). Standards for secure data sharing across organizations. Computer Standards & Interfaces, 29(1), 86-96. https://doi.org/10.1016/j.csi.2006.01.004

Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. Development Policy Review, 34(1), 135-174. https://doi.org/10.1111/dpr.12142

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E. J. (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in Methods and Practices in Psychological Science*, 1(1), 70-85. https://doi.org/10.1177/2515245917751886

Huang, R. L., Wang, X. K., Hou, W. H., Chen, Z. Y., Wang, Y. T., & Wang, J. Q. (2023). Two-way referral cooperative hospital selection with uncertain information: A two-sided matching decision-making approach. Computers & Industrial Engineering, 183, Article 109430. https://doi.org/10.1016/j.cie.2023.109430

Kass, N. E., Natowicz, M. R., Hull, R. C., Faden, R. R., Plantinga, L., Gostin, L. O., & Slutsman, J. (2003). The use of medical records in research: What do patients want? Journal of Law Medicine & Ethics, 31(3), 429-433. https://doi.org/10.1111/j.1748-720X.2003.tb00105.x

Kowalczyk, S., & Shankar, K. (2011). Data Sharing in the Sciences. Annual review of information science and technology, 45, 247-294. https://doi.org/10.1002/aris.2011.1440450113

Li, X., Lu, J. P., Hu, S., Cheng, K. K., De Maeseneer, J., Meng, Q. Y., . . . Hu, S. S. (2017). The primary health-care system in China. Lancet, 390(10112), 2584-2594. https://doi.org/10.1016/s0140-6736(17)33109-4

Li, X. J., & Cong, Y. L. (2024). Exploring barriers and ethical challenges to medical data sharing: perspectives from Chinese researchers. *Bmc Medical Ethics*, *25*(1), Article 132. https://doi.org/10.1186/s12910-024-01135-8

Palm, J., Södergren, K., & Bocken, N. (2019). The Role of Cities in the Sharing Economy: Exploring Modes of Governance in Urban Sharing Practices. Energies, 12(24), Article 4737. https://doi.org/10.3390/en12244737

Prajogo, D., Toy, J., Bhattacharya, A., Oke, A., & Cheng, T. C. E. (2018). The relationships between information management, process management and operational performance: Internal and external contexts. International Journal of Production Economics, 199, 95-103. https://doi.org/10.1016/j.ijpe.2018.02.019

Priego, L. P., Wareham, J., & Romasanta, A. K. S. (2022). The puzzle of sharing scientific data. Industry and Innovation, 29(2), 219-250. https://doi.org/10.1080/13662716.2022.2033178

Reddy, R. C., Mishra, D., Goyal, D. P., & Rana, N. P. (2024). A conceptual framework of barriers to data science implementation: a practitioners' guideline. *Benchmarking-an International Journal*, 31(10), 3459-3496. https://doi.org/10.1108/bij-03-2023-0160

Thoegersen, J. L., & Borlund, P. (2022). Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing. Journal of Documentation, 78(7), 1-17. https://doi.org/10.1108/jd-01-2021-0015

Weitzman, E. R., Kaci, L., & Mandl, K. D. (2010). Sharing Medical Data for Health Research: The Early Personal Health Record Experience. Journal of medical Internet research, 12(2), Article e14. https://doi.org/10.2196/jmir.1356

Xu, Q. L., Jiao, R. J., Yang, X., Helander, M., Khalid, H. M., & Opperud, A. (2009). An analytical Kano model for customer need analysis. Design Studies, 30(1), 87-110. https://doi.org/10.1016/j.destud.2008.07.001

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. Journal of the American society for information science and technology, 58(4), 479-493. https://doi.org/10.1002/asi.20508

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Collaborative human-AI risk annotation: co-annotating online incivility with CHAIRA

Jinkyung Katie Park, Rahul Dev Ellezhuthil, Pamela Wisniewski, and Vivek Singh

DOI: https://doi.org/10.47989/ir30iConf47146

Abstract

Introduction. Collaborative human-AI annotation is a promising approach for various tasks with large-scale and complex data. Tools and methods to support effective human-AI collaboration for data annotation are an important direction for research. In this paper, we present CHAIRA: a Collaborative Human-AI Risk Annotation tool that enables human and AI agents to collaboratively annotate online incivility.

Method. We leveraged Large Language Models (LLMs) to facilitate the interaction between human and AI annotators and examine four different prompting strategies. The developed CHAIRA system combines multiple prompting approaches with human-AI collaboration for online incivility data annotation.

Analysis. We evaluated CHAIRA on 457 user comments with ground truth labels based on the inter-rater agreement between human and AI coders.

Results. We found that the most collaborative prompt supported a high level of agreement between a human agent and AI, comparable to that of two human coders. While the AI missed some implicit incivility that human coders easily identified, it also spotted politically nuanced incivility that human coders overlooked.

Conclusions. Our study reveals the benefits and challenges of using AI agents for incivility annotation and provides design implications and best practices for human-AI collaboration in subjective data annotation.

Introduction

Online incivility refers to features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or the topic (Coe et al., 2014). Considering adverse effects (Gervais, 2015; Han et al., 2018), it is important to identify and moderate incivil comments on social media platforms as well as to understand the nature and characteristics of incivility (Coe et al., 2014; Jhaver et al., 2018; Matias, 2019; Oz et al., 2018), both of which require the annotation of digital trace data. This annotation process often includes crowdsourced workers (Hosseinmardi et al., 2015), a team of researchers and research assistants (Park et al., 2023b; Singh et al., 2017), and/or domain experts (Park et al., 2023a). It involves an intensive and collaborative process of training, consensus-building, and quality control among multiple coders; therefore, it can be costly and time-consuming, while still yielding uneven levels of inter-coder agreement (Coe et al., 2014; Rains et al., 2017). There is a growing need for innovative and efficient methods to support human coders in annotating large corpora of online data, which can have a significant methodological impact on information science research.

In this study, we explore the use of Large Language Model (LLM)-based Conversational Agents (CAs) as AI-based co-coders for annotating online incivility data. We focus on LLM-based CAs because they have shown promising results in text annotation tasks due to their accuracy and adaptability (Amin et al., 2023; Huang et al., 2023; Kuzman & Ljubešić, 2023; Liu et al., 2023; Zhang, et al., 2022a; Zhang et al., 2024). Moreover, LLMs can be adapted through finetuning or prompting for specific domains (Song et al., 2024), setting a new standard for what is achievable in natural language tasks. However, there are challenges in the use of LLM-based CAs in annotating textual data for more contextualized constructs (Amin et al., 2023; Huang et al., 2023), indicating the need for human-AI collaboration on subjective and contextualized annotation tasks.

In this paper, we present "CHAIRA: a Collaborative Human-AI Risk Annotation tool" that enables human and AI agents to co-annotate online incivility. We share early results from the design and implementation of a CHAIRA that we developed to interact with human coders and provide suggestions and explanations for annotating online incivility. Using 457 user comments with ground truth labels (e.g., civil vs. uncivil), we experiment with four types of prompting methods with different levels of information exchange between the human coder and CHAIRA. Using 10% of the data (n = 50), we established inter-rater agreement between the human coder and CHAIRA to observe how different types of promoting methods impact data annotation results. We analysed the conversation log between human coders and CHAIRA to have qualitative insights into how the quality of annotations changes with different prompting approaches. As such, we use a mixed methods approach to address the research questions:

> **RQ1:** how do different types of prompting methods influence the inter-rater reliability of human-AI collaborative data annotation results?
>
> **RQ2:** how do different types of prompting methods influence the quality and rationale for human-AI collaborative data annotation results?

By answering the above research questions, we address the overarching question: '*What is the optimal prompting approach and best practices to make the performance of human-AI collaboration similar to that of human-human collaboration*?' We found that CHAIRA's performance in terms of inter-coder agreement with human coders improved with more detailed prompts. The most advanced model, the Two-stage Few-shot Chain of Thought, nearly matched the agreement levels seen between two human coders reported in previous studies. While the AI agent missed some implicit incivility that human coders easily identified, it also spotted politically nuanced incivility that human coders overlooked. Our work provides design insights and best practices for human-AI collaboration in subjective data annotation tasks. It introduces a novel system for human-AI collaboration and applies different prompt engineering approaches to optimize incivility

annotation. These findings are applicable beyond online incivility scenarios, offering a path for scalable annotation in subjective or low-resource settings. As such, our work contributes to the iConference community by empirically demonstrating the potential of human-AI collaboration in the context of subjective digital trace data annotation. Particularly, we contribute to the iConference community's focus on addressing multifaceted dimensions of AI to foster a deeper understanding of their benefits, challenges, and broader implications.

Related work

Conversational agents as annotators

Conversational Agents (CAs) are systems enabled with the ability to interact with users using natural human dialogue (Rheu et al., 2021). After the recent release of various Large Language Model (LLM)-based Conversational Agents (CAs) (e.g., ChatGPT (OpenAI, 2022)), research communities are increasingly experimenting with data annotation tasks such as annotating political stance and sentiment of textual data (Amin et al., 2023; Kuzman & Ljubešić, 2023; Liu et al., 2023; Zhang et al., 2022a). Emerging literature suggests that LLM-based CAs can be useful for text classification tasks (e.g., Amin et al., 2023; Huang et al., 2023; Kuzman & Ljubešić, 2023; Liu et al., 2023; Zhang et al., 2022a; Zhang et al., 2024). For instance, Zhang et al. (2022) show that ChatGPT was able to annotate the political stance of the tweets with an average accuracy above 70. Moreover, LLMs can be adapted through finetuning or prompting for specific domains (Song et al., 2024), setting a new standard and expectations for what is achievable in natural language tasks. With proper fine-tuning, LLMs are known to even outperform crowdsourced annotators (Gilardi et al., 2023). As such, advances in LLMs such as GPT-4 showed a promising opportunity for data annotation at scale due to their ability to automate annotation tasks (Zhang et al., 2022b). However, there are challenges in the use of LLM-based CAs in annotating textual data for more contextualized constructs. For instance, Amin et al. (2023) showed that ChatGPT's accuracy for subjective tasks such as the five personality and suicide ideation classifications was lower than the baseline machine learning methods. As such, early empirical research demonstrated the limitations of subjective and contextualized annotation tasks entirely, indicating the need for human-AI collaboration on such tasks.

Human-AI collaboration on annotation

As LLMs-based conversational agents have shown the ability to interact with humans and work with examples in various domains (Kim et al., 2022; Lai et al., 2022; Mackeprang et al., 2019; Tang et al., 2024), researchers are exploring the potential of human-AI collaboration on various tasks such as online content moderation (Lai et al., 2022), thematic analysis of qualitative data (Jiang et al., 2021; Zhang et al., 2024), disease prevention (Lu & Peng, 2024), and crowdsourcing (Tamura et al., 2024). For instance, Zhang et al. (2024) explored the potential of LLM-based CAs as collaborative tools for qualitative data analysis and highlighted the efficiency of reducing time and labour for such analysis. Yet, their performance in collaborative co-annotation exercises for online risk where different facets of co-annotation are important is understudied. This gap is pertinent because co-annotation tasks need to support an interactive discussion to help generate a rationale for the various decisions, particularly in the context of highly contextualized online risk behaviour (Clay-Warnder, 2003) which can entail disagreement even among human coders. Yet, the disagreement must not lead to capitulation, instead, inspire better methods of automated analysis. Therefore, the combination of manual and automated content analysis is suggested as the gold standard for identifying subjective concepts such as online risk (Esau, 2022). In recent work, Wang et al. (2024) designed a multi-step human-LLM collaborative framework for data annotation tasks (e.g., natural language interface, stance detection, and hate speech detection) and found that when LLMs are incorrect in complex or domain-specific tasks, human annotations without any LLM assistance were the most accurate. Therefore, they suggested an iterative process of human-AI collaboration

in data annotation, with feedback from the human annotators used to improve the quality of LLMs annotation.

In this work, we explore the potential of using LLM-based CAs to assist human coders in annotating subjective, nuanced online conversations. We expect that LLM-based CAs can support high-quality annotations with explanations when provided with proper instructions and examples. This approach could enhance scalability and help capture nuances that human coders might miss due to cognitive limits. We experimented with four prompting methods to assess their impact on annotation results and how co-annotation can improve subjective data annotation. By examining agreements and disagreements between human coders and LLM-based CAs, we explored how their strengths and weaknesses can complement each other.

Methods

Design and system implementation of CHAIRA

CHAIRA is an online annotation tool that integrates an LLM-based conversational agent to support human-AI collaboration on online risk data annotation. Below, we describe how we designed and implemented CHAIRA in detail.

System implementation and dataset

We leveraged the GPT 3.5 Turbo (the underlying model for OpenAI's ChatGPT) as the LLM of choice due to its popularity and ease of use. We developed a custom annotation interface on top of OpenAI's API to support human-AI co-annotation. The interface was developed in React and deployed on AWS. We used S3 buckets to store the dataset and AWS Lambda to evaluate the dataset. We used the dataset collected in prior work in which researchers explored the effectiveness of embedding positive background images on online discussion forums in reducing online incivility (Park & Singh, 2022). The data comprised 457 comments collected from 105 users who participated in an online experiment and were annotated for online incivility by the two human coders. In the prior work (Park & Singh, 2022), researchers designed a codebook to annotate 457 use comments into civil vs. incivil and worked with a human coder (i.e., research assistant) to establish the reliability of the incivility coding scheme. After the researcher and the coder had multiple training sessions, 10% of the comments (n = 45) were coded to establish interrater reliability. The reported interrater reliability scores in prior work were 0.88 (percent agreement) and 0.76 (Cohen's Kappa). Once the interrater reliability was established, the researcher coded the rest of the data. In the final dataset, 55% were reported to be civil cases (n = 250) while 45% were incivil cases (n = 207) (Park & Singh, 2022).

In this work, we split the labelled dataset (457 use comments) into training (5%), validation (10%), and test (85%) datasets using a stratified random sampling approach. Consistent with common practices in human-human collaborative coding, less than 5% of data (20 user comments) were allocated to the training dataset to be used as examples and initial instructions (similar to training sessions for human-human coding). Then each prompt was passed to 50 samples (approx. 10% of data) allocated to the validation dataset to evaluate the inter-rater reliability between the human coder and CHAIRA. The rest of the comments (387 comments) were allocated to the test dataset to evaluate human-AI agreement on final online risk data annotation results.

Following the common human-human coding practices, we split our training, validation, and test datasets to be independent of each other. For instance, samples from either the validation or test split were not mixed with the training split. In addition, only comments from the training set can be added as examples in the prompts to fine-tune CHAIRA. Similarly, samples from the validation set can be used to interact with CHAIRA but cannot be added as examples in a prompt. Samples from the test split cannot be loaded into the prompt.

Design of web interface

The web interface of CHAIRA (see Figure 1) provides an overview of the layout for a human coder to interact with CHAIRA and design prompts. The left side of the interface shows a list of different prompts created for online risk annotation tasks in black labels. Once the human coder clicks a specific prompt, the label of the prompt turns blue to indicate that it is an active prompt. The right side of the interface shows the chosen prompt, comment data, conversation log between the human coder and AI agent, and inter-rater agreement. Below, we zoom into the major components of the CHAIRA interface to describe how each component was designed to facilitate human-AI collaboration on online risk data annotation tasks.

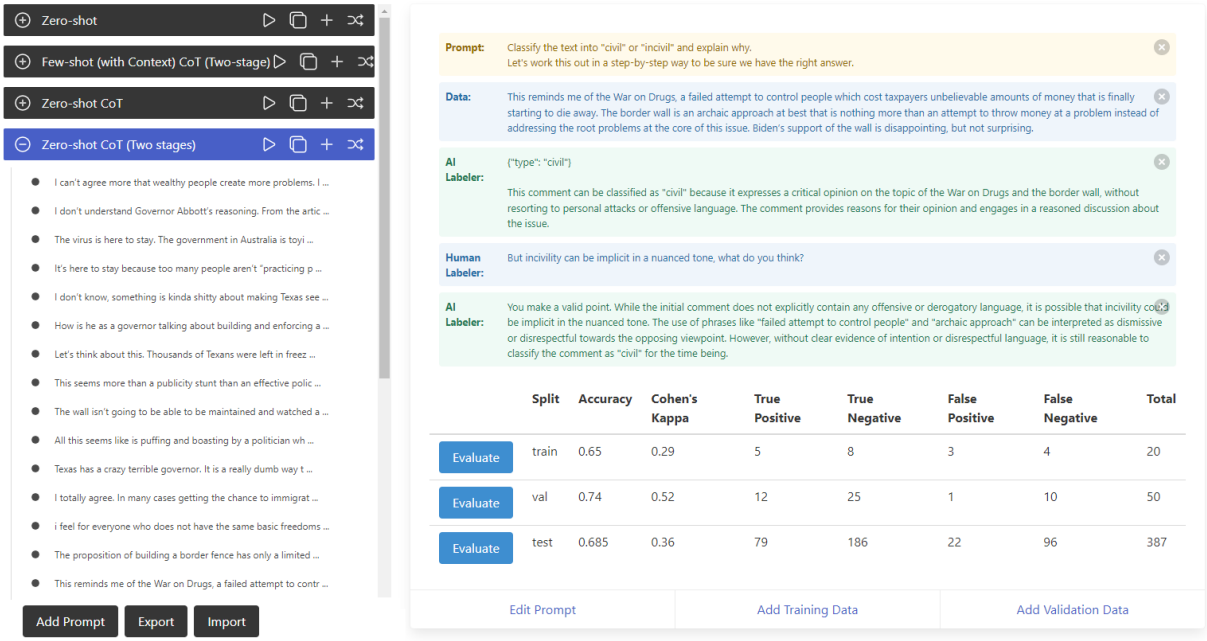


Figure 1. Overview of CHAIRA web interface. A list of designed prompts is shown on the left side, while the prompt, conversation log, and evaluation metrics/results are shown on the right side

Creating prompts: On the left side of the interface, the human coders can create new prompts to interact with an AI agent by clicking 'Add Prompt' (Figure 2). Once clicking the add prompt, a new text box appears where human coders can add the name of a certain prompt in the 'Prompt label' box and add the content of the prompt in the 'Prompt text' box to create a new prompt. Once new prompts are created, the human coders can add sample comments to test with the new prompts. A double arrow icon (on the left side of the red box in Figure 2 helps human coders to randomly sample comments from the training data. The user comments are added within the threads under each prompt with bullet points. Beyond the labelled dataset (457 user comments), the human coders can manually add new comments to label incivility using the same prompt by clicking a plus icon in the middle of the red box in Figure 2. Yet, these comments are not included when evaluating the prompt for inter-rater agreement with a human coder. Human coders can create copies of existing prompts by clicking the double square icon on the left in the red box in Figure 2. We chose icons for the above three features as there is limited space allocated in the prompt label (Figure 1). The 'Export' feature helps human coders download the prompt and conversation log data in a JSON file format, while the 'Import' feature helps the opposite, uploading JSON files to the interface.

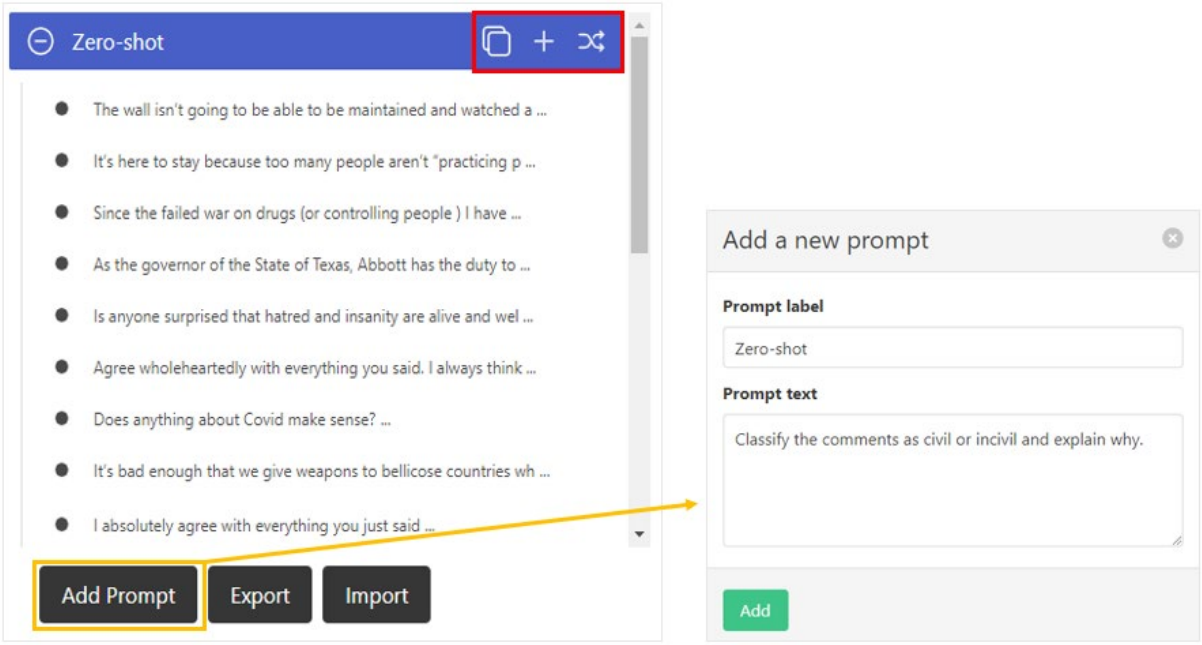


Figure 2. Features to create and manage prompts to interact with CHAIRA

Evaluating inter-rater agreement: Inter-rater agreement between the human coder and the AI agent can be assessed using the 'Evaluate' buttons (Figure 3). To report inter-rater agreement between the human coder and CHAIRA, we used percent agreement and Cohen's Kappa, following the practices in the literature on qualitative content analysis (McDonald et al., 2019; Tinsley & Weiss, 2000). 'Add Training Data' loads all 20 user comment data from the training dataset to be used as examples and initial instructions. Once the human coders click the button, the interface creates a thread under the prompt to display all 20 user comment data (see the left side of Figure 1). 'Add Validation Data' loads all 50 samples allocated in the validation dataset to establish the inter-rater agreement between the human coders and CHAIRA. After looking at the evaluation results, the human coder can edit the prompt by clicking the 'Edit Prompt' button. The interface can evaluate multiple prompts at the same time, which supports its scalability.

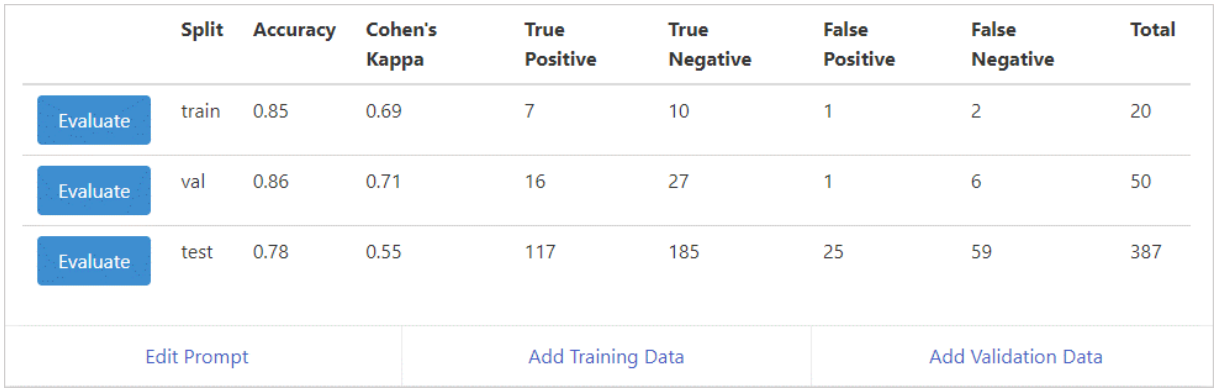


Figure 3. Features to assess inter-rater reliability between the human coders and CHAIRA

Human-AI interaction: Once the interface loads user comment data from the training set, the human coders can have interactive conversations with the AI agent on the right side of the interface. Once the human coder clicks a certain user comment under the prompt, the interface shows the prompt, comment data, and response from CHAIRA (Figure 4).

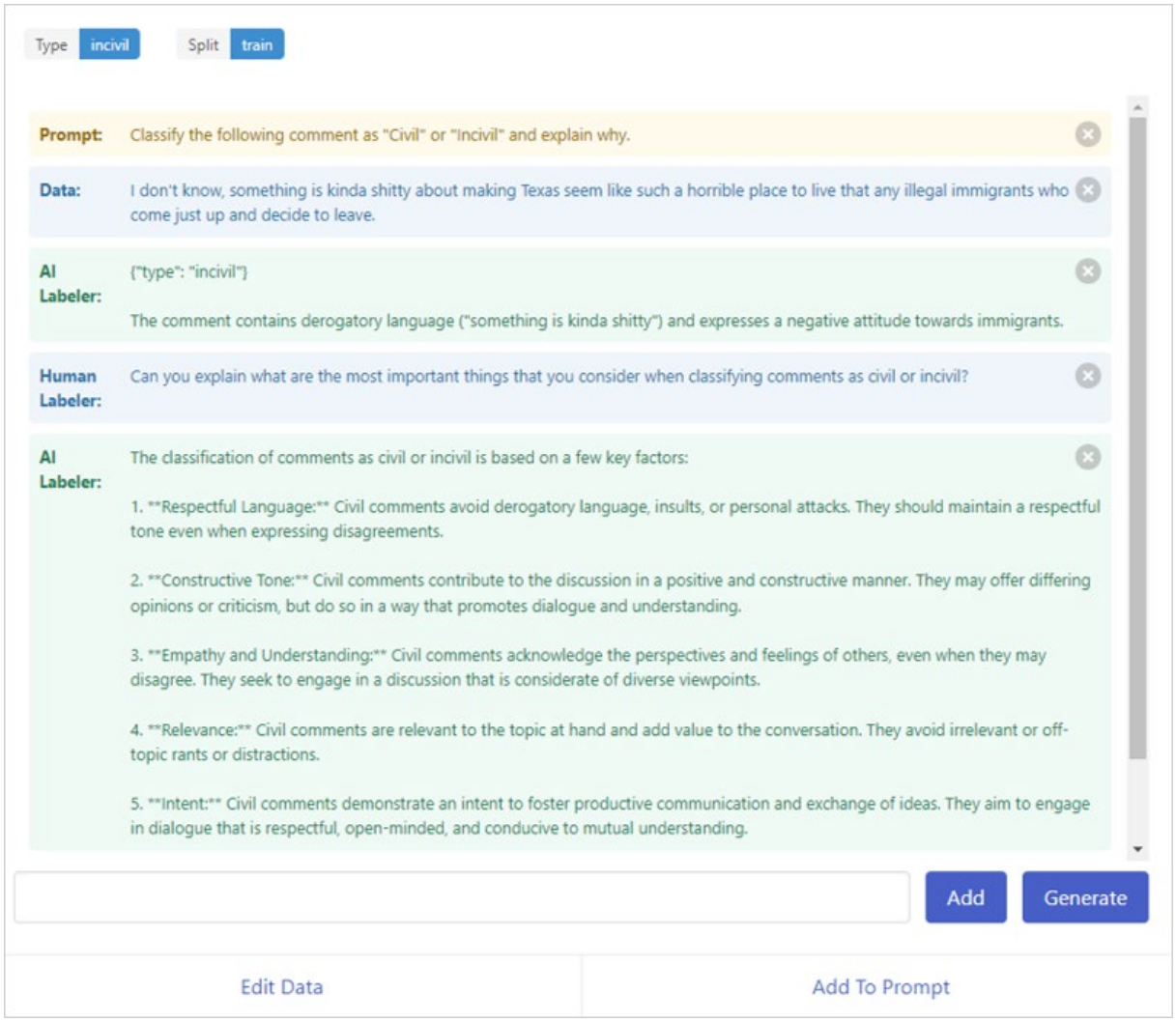


Figure 4. Features to support interactive communication between the human coders and CHAIRA

The 'Type' icon on the top left indicates an incivility label annotated by the human coder in prior work. The 'Split' icon indicates to which dataset the user comment data belongs. As seen from Figure 4, 'incivil' and 'train' means that the user comment data came from the training dataset and was annotated as incivil by the human coder. Right below these icons, the text in the prompt is shown as brown text with the label 'Prompt' on a yellow background. User comment data to annotate for incivility comes next as blue text with the label 'Data' in a light-blue background. Then the incivility labels and the rationales for the decision generated by CHAIRA follow as green text with the label 'AI Labeller' in a light-green background. The three components above are automatically generated when evaluating the inter-rater reliability for each prompt.

After reviewing the initial response generated by CHAIRA, the human coders can start a conversation by adding queries in the textbox and clicking the 'Add' button next to the textbox. Then, when the human coders click the 'Generate' button, CHAIRA generates responses to the given queries. The queries asked by the human coders appear with the label 'Human Labeller' in a light-blue background, while the answers generated by CHAIRA appear with the label 'AI Labeller' in a light-green background (Figure 4). We designed text generated by CHAIRA to appear as green

text in light-green backgrounds while text submitted by the human coders to appear as blue text in light-blue backgrounds to help human coders distinguish text generated by the two parties.

Through these interactive conversations, additional instructions and examples are exchanged between the two. Once the human coder decides that a reasonable agreement was achieved, a conversation log between the human coder and CHAIRA can be added as a prompt by clicking 'Add To Prompt' (Two-stage prompt in the next section). Following the common practices in human-human collaborative coding in which approximately 5-10% of the data is used for training and consensus building, we designed only conversation around 20 user comments in the training dataset to be added as prompts using the 'Add To Prompt' feature. Finally, the 'Edit Data' feature helps the human coders edit the user comment data.

Prompt engineering approaches

We conducted experiments with four different prompting engineering approaches: zero-shot, definition, few-shot, and two-stage few-shot Chain-of-Thought (CoT). We designed four different prompting approaches with varying levels of interaction between the human coder and the LLM-based agent. We used the same coding scheme as applied in the previous literature (Park & Singh, 2022) to design the prompts. In our prompts, incivility was defined as 'the feature of discussion that conveys an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topic' (Coe et al., 2014), with six different categories: name-calling, aspersion, lying, vulgarity, pejorative for speech, and others (Table 1). The reported inter-rater agreement between the two human coders was 0.88 (percent agreement) and 0.76 (Cohen's Kappa) (Park & Singh, 2022).

Category	Description	Example
Name-calling	Mean-spirited or disparaging words directed at a person or group of people	'*At least the morons in the state capital no longer have control of this process!*'
Aspersion	Mean-spirited or disparaging words directed at an idea, plan, policy, or behaviour. An aspersion may be both explicit and implicit	'*It beckons the memories of Trump's silly border wall, and the incredible waste of resources that was*'
Lying	Stating or implying that an idea, plan, policy, or public figure was disingenuous	'*Government is wrong, is corrupt, is lying, is deceiving the people, and is violating our constitution*'
Vulgarity	Using profanity of language that would not be considered proper in professional discourse	'*Am I possibly the only person here who thinks this shit is funny as hell?*'
Pejorative for speech	Disparaging remark about the way in which a person communicates	'*Quit crying over the spilled milk of*'
Others	All comments that may be deemed incivil, but do not fall into any of the previous categories of incivility	'*Hahahahahahahahahahaha,, really crack me open this one*'

Table 1. Definition and examples of types of incivility (Park & Singh, 2022)

Zero-shot prompting: In zero-shot prompting, the model is only given a simple instruction describing the task. This method is considered convenient and has the potential for robustness (Brown et al., 2020). The instruction used in the zero-shot prompt is as follows: 'Classify the text into 'civil' or 'incivil' and explain why'.

Definition prompting: In definition prompting, along with the instruction of classifying a comment as 'civil' or 'incivil,' we provided the definition of incivility and brief descriptions of six categories of incivility (see Table 1).

Few-shot prompting: With few-shot prompting, models are given a few demonstrations of the task (Brown et al., 2020), in our case, examples of incivility. With this approach, we provided the model with the definition of incivility, descriptions of the six categories of incivility, examples of the six categories of incivility (Table 1), and the instructions for the task.

Two-stage few-shot chain-of-thought: Finally, we used a two-stage few-shot chain-of-thought (CoT), a few-shot-based prompting for the chain of thought reasoning. Chain-of-thought (CoT) prompting (Wei et al., 2022) modifies the answers in few-shot examples to step-by-step answers by adding an instruction such as 'Let's think step by step' to the original prompt to elicit reasoning in LLMs (Kojima et al., 2022; Wei et al., 2022). The first prompt is the reasoning extraction where we used CoT to elicit reasoning from LLM. The second prompt consists of the first prompt and the answers generated from the first prompt.

Figure 5 shows a visualization of the two-stage few-shot CoT approach applied. In the first prompt, we provided the same instructions as in few-shot prompting and added the final line that says, 'Let's work this out in a step-by-step way to be sure we have the right answer' suggested by Zhou et al. (2022). Next, we reviewed the responses generated by CHAIRA from the first prompt and did the error analysis. We chose one example of false positive cases (i.e., CHAIRA output = civil, human ground truth = incivil) and pointed the model to recognize what it is missing in its answers (i.e., implicit aspersion). Once CHAIRA generated the responses that matched with human responses, we added the conversation log (Figure 6 in Appendix A) to the prompt. A summary of the four prompting approaches we used in this case study is presented in Table 2.

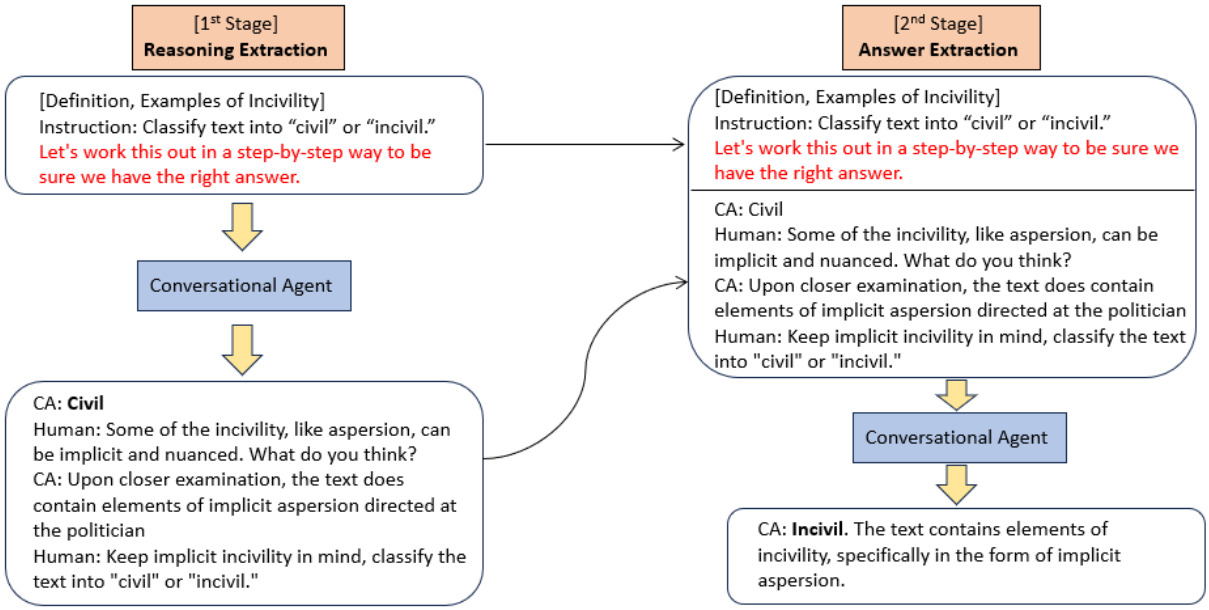


Figure 5. Pipeline of two-stage chain of thought prompting. Human feedback on phase 1 and CA responses are prepended to the input for phase 2 sent to CA

Results

Our methodology, system implementation, and prompting strategies showed that practical systems for human-AI collaboration in online risk annotation are feasible. To investigate RQ1, we measured the inter-coder agreement between the human coder and the AI agent. The agreement increased with the amount of detail in the prompts. The inter-rater agreement increased with additional information given in the prompts. Two-stage CoT yielded the highest performance (Cohen's Kappa = 0.71), yet, lower than the levels observed for the baseline, two human coders (Cohen's Kappa = 0.76) (Park & Singh, 2022) (see Table 2).

	Prompt				Performance	
	Instruction	Definition	Example	Conversation log	Percent agreement	Cohen's Kappa
Zero-shot	X				0.66	0.26
Definition	X	X			0.72	0.48
Few-shot	X	X	X		0.78	0.54
Two-stage CoT	X	X	X	X	0.86	0.71
Baseline	X	X	X	X	0.88	0.76

Table 2. Summary and performance of the four prompting approaches

To further understand the impact of different prompting methods on the annotation quality and rationale (RQ2), we discuss some of the common themes observed in the logs of interaction between the human coder and CHAIRA.

CHAIRA performed better with more human-AI interaction in the prompts, particularly for annotating implicit incivility

Overall, CHAIRA did a better job at annotating explicit incivility and explaining the reasoning behind them with more information provided in the prompts. For instance, even with the zero-shot prompting approach, CHAIRA pointed to the exact incivil expressions and recognized that such expressions reflect a negative attitude towards immigrants. In addition, without any information about incivility, CHAIRA was able to note the use of more nuanced incivility, such as sarcasm, when nuanced expressions are combined with explicit personal attacks and insults (examples in Figure 7 in Appendix).

At the same time, we observed some common issues throughout all the prompting approaches: CHAIRA did not recognize implicit and nuanced incivil expressions in the texts, even with the information 'an aspersion may be both explicit and implicit' given in the prompt. Therefore, when designing the two-stage few-shot CoT prompt, we reminded CHAIRA that aspersion can be implicit and nuanced through interactive conversation. Only after the interactive conversation between the human coder and CHAIRA about implicit aspersion was added to the prompt (i.e., the two-stage few-shot prompting approach), CHAIRA started to distinguish implicit and nuanced yet incivil expressions (examples in Figure 8 in Appendix), which were frequent in our dataset.

The output label remained the same, but the reasons changed with different prompts

We observed that while the output label remained the same the reasons changed with more information in the prompts. For instance, with the zero-shot prompt, CHAIRA mainly focused on the use of language and tone of the text, whereas with the definition prompt, CHAIRA considered whether the text fell under any of the six categories of incivility (Figure 7). The rationales provided were similar for definition and few-shot prompting approaches. A similar trend was observed for incivil cases, where CHAIRA provided specific reasons by pointing to the specific category (i.e., lying) and the context of why it falls under the category with the definition and few-shot prompting approaches, while CHAIRA provided more generic reasons (i.e., use of personal attacks or offensive language) with the zero-shot approach.

Overall, CHAIRA could provide human coders with helpful context on US politics, immigration policy debates, and incivil expressions that may be overlooked. For example, the text '*And around Nancy's wall on Capitol Hill. Make that wall to keep them in*' was annotated as civil by a human coder, as it did not contain explicit or implicit incivility. However, with zero-shot and definition prompts, CHAIRA recognized 'them' as immigrants, and the comment about building a wall to 'keep them in' was interpreted as disrespectful. Using a two-stage few-shot CoT approach, CHAIRA's

rationale became even more detailed: 'The comment mocks Nancy Pelosi and implies hostility toward immigrants.' The human coder missed that 'Nancy' referred to a political figure and failed to recognize the disparaging tone. Although another coder might have annotated it differently, the above example shows that human coders have limits in awareness and cognitive capacity and AI can complement such limitations. As such, with the highest level of human-AI interaction (two-stage CoT), CHAIRA effectively discerned both the tone and the target of incivility, providing crucial context in annotating political incivility.

Yet, sometimes, CHAIRA did not fully understand the information in the prompts and text to annotate

Unlike human conders, we observed some cases where CHAIRA did not accurately catch the textual information given in the prompts. For instance, CHAIRA lacked an understanding of the description and example of the 'Pejorative for Speech' category. We found some of the responses that CHAIRA mistakenly understood the pejorative for speech as the use of a sarcastic tone in the text, as opposed to its actual concept, a disparaging remark about how a person communicates. Similarly, CHAIRA sometimes lacked an understanding of the given text compared to human coders, particularly for short texts. For instance, with the definition prompting approach, CHAIRA struggled to annotate a short text such as 'What is your solution?' hence, annotated it as 'unclear.' With the few-shot prompt, CHAIRA annotated the same text as incivil because the text contains examples of name-calling and aspersions. However, the incivil expressions referred to in this response were from the examples given in the prompt (i.e., instructions), not from the given text to annotate. Overall, sometimes CHAIRA could not understand the information in the prompts or texts to annotate and hence, was unable to annotate the text appropriately.

Implications for building human-AI collaborative annotation systems

The above-mentioned results yielded implications for the future design of human-AI collaborative annotation systems. Below, we discuss the design implications for using AI-based CAs to best support the co-annotation of online risk data.

Reasoning and domain knowledge provided by CAs are valuable resources for co-annotation workflows

We observed that CHAIRA was good at providing reasons for its annotation results. Therefore, the benefit of CHAIRA lies in the interactive nature of the annotation process, which provides partial explainability and is one of the important aspects of human-centred AI-based systems (Minh et al., 2022; Vilone & Longo, 2021). With the most sophisticated prompting approach we had (Few-shot CoT), CHAIRA informed the human coder with a broad knowledge and context of the given text and convinced the human coder to change their mind in multiple instances. Therefore, one of the strengths of LLM-based CAs is the ability to provide relevant information (presumably) trained on the entirety of online data, as opposed to human coding which requires extensive training or domain-specific knowledge. This shows the potential strengths and values in human-AI co-annotation, particularly in risk scenarios that require domain-specific knowledge. In addition, the initial reasoning and domain knowledge provided by CAs can be used to inform human coders further on how to design better CA models to support the co-annotation of online risk.

Two-way interaction between human coders and CAs is a key to good co-annotation results

A major benefit of co-coding with CHAIRA was its ability to scale the data annotation with a high degree of inter-rater agreement. To be able to do so, we carefully reviewed the incivility labels where CHAIRA and the human coder disagreed and had two-way conversations with CHAIRA to further understand their reasoning. This two-way interaction in the co-annotation process was useful because, despite access to a large corpus of knowledge, CHAIRA also tended to make some

simple mistakes that were quite easy for human coders to spot. For instance, during the interaction with CHAIRA, we realized that CHAIRA could miss the previous conversation and hence, we needed to remind it about our conversation for further annotation tasks, particularly about nuanced incivility (e.g., implicit aspersion). Therefore, we added the instruction '*keep implicit incivility in mind*,' in our Two-stage CoT prompt. Sometimes, we pointed to the exact incivil expressions in the text that contain implicit aspersion to remind the concept (e.g., Don't you think this expression of 'he's hoping to stir up the same frenzy and ride that wave?' could be implicitly incivil?). Then the CHAIRA re-evaluated the text and corrected their answers. As such, interactive communication between the human coders and the AI agent is one of the key elements in improving the risk annotations generated by the AI agent.

Providing clear examples in carefully designed prompts considering how LLMs process human language is important

Providing clear examples of risk cases and descriptions of risk types is crucial when designing human-AI co-annotator models. In our few-shot prompt, we explained that aspersion can be both implicit and explicit, yet CHAIRA failed to recognize implicit aspersion until we guided it through two-way communication. This could be due to confusion caused by slight differences between the risk descriptions and examples provided. For instance, CHAIRA may have interpreted the explicit nature of the 'silly border wall' in the aspersion example and missed the implicit aspect described. Therefore, selecting the right examples and crafting clear descriptions of constructs is critical when working with LLM-based CAs to annotate subjective concepts like online risk. In addition, since CAs generate responses by tokenizing input (OpenAI, 2024b), even minor textual changes such as punctuation can affect performance. This can limit the CA's ability to understand text containing abbreviations or spelling variations, which are common in online risk data (Sadeque et al., 2019). Therefore, designing prompts with careful consideration of how LLMs process natural language is essential to building effective collaborative systems for annotating contextualized online risk data.

Limitations and future directions

In consonance with data ethics, we used de-identified data to develop our annotation tool. We used OpenAI API in the backend as its security policy stipulates that data submitted through the API is not used to train OpenAI models (OpenAI, 2024a). However, future work should also consider building LLM-based CAs with private servers so that the training dataset is not shared via the web. Another limitation of this collaborative annotation is variations in LLM responses. Finally, we acknowledge that inductive approaches (e.g., thematic analysis, grounded theory approach), important approaches to building patterns and themes in qualitative work, were not explored in this study. Future work can explore the potential of LLM-based CAs for more inductive analysis that requires an in-depth understanding of the subtleties and complexities of qualitative data. Moving forward, we aim to experiment with our approaches with diverse types of online risk data at scale to gain deeper insights into collaborative annotation between human and LLM-based conversational agent systems.

Conclusion

In this study, we built systems to support human-AI collaborative data annotation tasks and explored the potential benefits and challenges of human-AI collaborative annotation of highly subjective and contextualized online incivility data. The AI missed some implicit risks that human coders easily spotted, conversely, it spotted politically nuanced incivility that human coders overlooked. The design implications and best practices derived from this work can serve as a steppingstone for future research considering similar methods. Our work suggests a path toward combining the relative strengths of humans and AI for scalable data annotation, especially in sensitive or low-resource settings.

About the authors

Jinkyung Katie Park is an Assistant Professor in the School of Computing at Clemson University, Clemson, USA. She received her Ph.D. from Rutgers University, and her research focuses on Human-Computer Interaction to promote the online safety of vulnerable populations. She can be contacted at jinkyup@clemson.edu

Rahul Dev Ellezhuthil is a Data Scientist who received a master's degree in Computer Science from Rutgers University. He can be contacted at rahul.e.dev@gmail.com

Pamela Wisniewski is an Associate professor in the Department of Computer Science at Vanderbilt University, Nashville, USA. Her work lies at the intersection of Human-Computer Interaction, Social Computing, and Privacy. She can be contacted at pamela.wisniewski@vanderbilt.edu

Vivek Singh is an Associate professor in the School of Communication and Information at Rutgers University, New Brunswick, USA. He designs AI systems that are responsive to human values and needs. He can be contacted at vivek.k.singh@rutgers.edu

References

Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. IEEE Intelligent Systems, 38(2), 15-23. https://doi.org/10.1109/MIS.2023.3254179

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Clay-Warner, J. (2003). The context of sexual violence: Situational predictors of self-protective actions. Violence and victims, 18(5), 543. https://doi.org/10.1891/088667003780928099

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. Journal of communication, 64(4), 658-679. https://doi.org/10.1111/jcom.12104

Esau, K. (2022). Content analysis in the research field of incivility and hate speech in online communication. In Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft–Standardized Content Analysis in Communication Research: Ein Handbuch-A Handbook (pp. 451-461). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-36179-2_38

Gervais, B. T. (2015). Incivility online: Affective and behavioural reactions to uncivil political posts in a web-based experiment. Journal of Information Technology & Politics, 12(2), 167-185. https://doi.org/10.1080/19331681.2014.997416

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Han, S. H., Brazeal, L. M., & Pennington, N. (2018). Is civility contagious? Examining the impact of modelling in online political discussions. Social Media+ Society, 4(3), 2056305118793404. https://doi.org/10.1177/2056305118793404

Hosseinmardi, H., Mattson, S. A., Ibn Rafiq, R., Han, R., Lv, Q., & Mishra, S. (2015). Analysing labelled cyberbullying incidents on the Instagram social network. In Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7 (pp. 49-66). Springer International Publishing. https://doi.org/10.1007/978-3-319-27433-1_4

Huang, F., Kwak, H., & An, J. (2023, April). Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In Companion proceedings of the ACM web conference 2023 (pp. 294-297). https://doi.org/10.1145/3543873.3587368

Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. ACM Transactions on Computer-Human Interaction (TOCHI), 25(2), 1-33. https://doi.org/10.1145/3185593

Jiang, J. A., Wade, K., Fiesler, C., & Brubaker, J. R. (2021). Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-23. https://doi.org/10.1145/3449168

Kim, T. S., Choi, D., Choi, Y., & Kim, J. (2022, April). Stylette: Styling the web with natural language. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-17). https://doi.org/10.1145/3491102.3501931

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.

Kuzman, T., Ljubešić, N. (2023). Automatic genre identification: a survey. Lang Resources & Evaluation, 1-34. https://doi.org/10.1007/s10579-023-09695-8

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022, April). Human-ai collaboration via conditional delegation: A case study of content moderation. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-18). https://doi.org/10.1145/3491102.3501999

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology, 100017. https://doi.org/10.1016/j.metrad.2023.100017

Lu, Q., & Peng, X. (2024, April). Differences in Knowledge Adoption Among Task Types in Human-AI Collaboration Under the Chronic Disease Prevention Scenario. In International Conference on Information (pp. 213-231). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57867-0_16

Mackeprang, M., Müller-Birn, C., & Stauss, M. T. (2019). Discovering the sweet spot of human-computer configurations: A case study in information extraction. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-30. https://doi.org/10.1145/3359297

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. Proceedings of the National Academy of Sciences, 116(20), 9785-9789. https://doi.org/10.1073/pnas.1813486116

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. Proceedings of the ACM on human-computer interaction, 3(CSCW), 1-23. https://doi.org/10.1145/3359174

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 1-66. https://doi.org/10.1007/s10462-021-10088-y

Open AI. (2022). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

Open AI. (2024a). *Security & privacy*. https://openai.com/security-and-privacy/

Open AI (2024b). *What are tokens and how to count them?* https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. New media & society, 20(9), 3400-3419. https://doi.org/10.1177/1461444817749516

Park, J., Ellezhuthil, R. D., Isaac, J., Mergerson, C., Feldman, L., & Singh, V. (2023a). Misinformation detection algorithms and fairness across political ideologies: The impact of article level labelling. In Proceedings of the 15th ACM Web Science Conference 2023 (pp. 107-116). https://doi.org/10.1145/3578503.3583617

Park, J., Gracie, J., Alsoubai, A., Stringhini, G., Singh, V., & Wisniewski, P. (2023, April). Towards automated detection of risky images shared by youth on social media. In Companion Proceedings of the ACM Web Conference 2023 (pp. 1348-1357). https://doi.org/10.1145/3543873.3587607

Park, J., & Singh, V. K. (2022). How Background Images Impact Online Incivility. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), 1-23. https://doi.org/10.1145/3555545

Rains, S. A., Kenski, K., Coe, K., & Harwood, J. (2017). Incivility and political identity on the Internet: Intergroup factors as predictors of incivility in discussions of news online. Journal of Computer-Mediated Communication, 22(4), 163-178. https://doi.org/10.1111/jcc4.12191

Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2021). Systematic review: Trust-building factors and implications for conversational agent design. International Journal of Human–Computer Interaction, 37(1), 81-96. https://doi.org/10.1080/10447318.2020.1807710

Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2019, June). Incivility detection in online comments. In Proceedings of the eighth joint conference on lexical and computational semantics (* SEM 2019) (pp. 283-291). http://doi.org/10.18653/v1/S19-1031

Singh, V. K., Ghosh, S., & Jose, C. (2017, May). Toward multimodal cyberbullying detection. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2090-2099). https://doi.org/10.1145/3027063.3053169

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., & Wang, H. (2024, March). Preference ranking optimization for human alignment. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 18990-18998). https://doi.org/10.1609/aaai.v38i17.29865

Tamura, T., Ito, H., Oyama, S., & Morishima, A. (2024, April). Influence of AI's Uncertainty in the Dawid-Skene Aggregation for Human-AI Crowdsourcing. In International Conference on Information (pp. 232-247). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57867-0_17

Tang, Y., Chang, C. M., & Yang, X. (2024, March). PDFChatAnnotator: A Human-LLM Collaborative Multi-Modal Data Annotation Tool for PDF-Format Catalogs. In Proceedings of the 29th International Conference on Intelligent User Interfaces (pp. 419-430). https://doi.org/10.1145/3640543.3645174

Tinsley, H. E., & Weiss, D. J. (2000). Interrater reliability and agreement. In Handbook of applied multivariate statistics and mathematical modelling (pp. 95-124). Academic Press. https://doi.org/10.1016/B978-012691360-6/50005-7

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89-106. https://doi.org/10.1016/j.inffus.2021.05.009

Wang, X., Kim, H., Rahman, S., Mitra, K., & Miao, Z. (2024, May). Human-LLM collaborative annotation through effective verification of LLM labels. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-21). https://doi.org/10.1145/3613904.3641960

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

Zhang, B., Ding, D., Jing, L., Dai, G., & Yin, N. (2022a). How would stance detection techniques evolve after the launch of chatgpt?. arXiv preprint arXiv:2212.14548. https://doi.org/10.48550/arXiv.2212.14548

Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Kim, C., ... & Cai, J. (2024). When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding. arXiv preprint arXiv:2407.14925. https://doi.org/10.48550/arXiv.2407.14925

Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022b). Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493. https://doi.org/10.48550/arXiv.2210.03493

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910. https://doi.org/10.48550/arXiv.2211.01910

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Appendix

Data	All this seems like is puffing and boasting by a politician who is coming up for reelection and has a prospective presidential run in the future. Just trying to get a check mark next to his name for being strong on border and immigration without actually doing much to change, fix, or address the problem at all. All the while crying out at every chance you get how it is your oppositions fault. The real problem here is the consequences of there actions. The U.S citizens whose land is going to be taken from them in imminent domain. Those who are now completely blocked from access to the water way. The years and years of lawsuits these people are going through.
AI labeler	Type: Civil. Explanation: The text expresses criticism and frustration towards a politician and their actions, but it does not contain any explicit name-calling, aspersions, lying, vulgarity, pejorative for speech, or other uncivil language. It focuses on the consequences and impact of the politician's actions, which can be seen as a legitimate critique. Therefore, it can be classified as civil.
Human labeler	Some of the incivility, like aspersion, can be implicit and nuanced. What do you think?
AI labeler	Type: Incivil. Upon closer examination, the text does contain elements of implicit aspersion directed at the politician. The language used suggests that the politician is simply "puffing and boasting" without actually taking meaningful action to address the border and immigration problem. The text also implies that the politician is blaming their opposition for the issue without taking responsibility themselves. These implicit aspersions contribute to an overall tone of disrespect towards the politician. Therefore, the text can be classified as incivil.
Human labeler	Keep implicit incivility in mind, classify the text into "civil" or "incivil."

Figure 6. Conversation log that was added to the two-stage few-shot CoT prompt

Text	What does combating global climate crisis even mean? Why does anyone thing politicians have answers. They push this agenda for additional tax money. Time magazine said there was a mini ice age coming in the 1970's. Then Al Gore (internet inventor) said we had global warming that would drown all of Florida and half of New York right now. Now we have climate change. Yes we do, everything changes. Everything has a cycle. Nothing is constant. If the climate didn't change government would raise taxes telling us they could change the stagnant climate. Wake up people…"
Zero-shot response	This comment is incivil because it resorts to **personal attacks and insults** toward politicians, insinuates ulterior motives for advocating climate action, and dismisses the scientific consensus on climate change. It also uses **sarcasm and an accusatory tone** towards the readers, which further contributes to its incivility.
Definition response	This comment falls under the category of **"lying"** because it implies that politicians are pushing the climate agenda solely for additional tax money, without providing any evidence or factual basis for this claim. It also contains **sarcastic remarks and dismissive language, which contributes to the disrespectful tone of the comment.**

Figure 7. Comparison between the responses from the incivil case generated with zero-shot vs. definition prompting approaches

Text	It's bad enough that we give weapons to bellicose countries which use those weapons to attack weaker ones, and then spend even more money sending those countries aid. There's a sick lack of empathy for people right across the border, as you said. So many people here are incapable of putting themselves in another's shoes, unable to imagine being born into less fortunate circumstances, and having to do whatever it takes to survive and give your children a better chance in a safer country. But then, once they need help in Texas, they want to be bailed out and point the blame at anyone else.
Zero-shot response	This comment expresses a strong opinion and criticism towards the government and the lack of empathy for people in other countries. However, it does so in a respectful and rational manner, addressing a common issue of hypocrisy that some individuals may have. While it may be a controversial statement, **it does not contain any personal attacks or offensive language, making it a civil comment.**
Few-shot response	The comment expresses a strong opinion about the topic and may be seen as critical or unkind towards the people who hold a certain viewpoint, but it **does not contain any explicit name-calling, aspersions, lies, vulgarity, pejorative for speech, or other forms of incivility**. The language used is generally respectful, even though it may convey frustration or disagreement.
Two-stage CoT response	The text contains elements of incivility. **It includes implicit aspersions towards certain groups of people who lack empathy and blame others when in need.** The language used, such as "incapable," "unable," and "point the blame," conveys a disrespectful tone towards these individuals. Additionally, the text exhibits a divisive attitude by highlighting the perceived hypocrisy of those in Texas. Therefore, the text can be classified as incivil.

Figure 8. Comparison among responses for the implicit incivil case with different prompting approaches

Information Research - Vol. 30 No. iConf (2025)

Deskilling and upskilling with AI systems

Kevin Crowston and Francesco Bolici

DOI: https://doi.org/10.47989/ir30iConf47143

Abstract

Introduction. Deskilling is a long-standing prediction of the use of information technology, raised anew by the increased capabilities of AI (AI) systems. A review of studies of AI applications suggests that deskilling (or levelling of ability) is a common outcome, but systems can also require new skills, i.e., upskilling.

Method. To identify which settings are more likely to yield deskilling vs. upskilling, we propose a model of a human interacting with an AI system for a task. The model highlights the possibility for a worker to develop and exhibit (or not) skills in prompting for, and evaluation and editing of system output, thus yielding upskilling or deskilling.

Findings. We illustrate these model-predicted effects on work with examples of current studies of AI-based systems.

Conclusions. We discuss organizational implications of systems that deskill or upskill workers and suggest future research directions.

Introduction

The increased capability of modern artificial intelligence (AI) systems, generative AI in particular, has increased concerns about their impact. We define AI as '*systems that build on machine learning, computation, and statistical techniques, as well as rely on large data sets to generate responses, classifications, or dynamic predictions*' (Faraj, Pachidi, & Sayegh, 2018, p. 62). In this paper we focus on a long-standing concern about the impact of automation, namely deskilling, meaning that the work left for the humans requires a lower level of skill than the original job. AI raises the question of deskilling anew since as a general-purpose technology that could impact more kinds of work (Sison, Daza, Gozalo-Brizuela, & Garrido-Merchán, 2023). Consistent with the fear of deskilling, many AI applications are described as having a levelling effect, meaning that they help novices more than experts (i.e., levelling ability), which we interpret as deskilling. For instance, Brynjolfsson, Li, and Raymond (2023) found that a chatbot to support customer service workers enabled less experienced operators work at the level of more experienced one.

However, as Crowston and Bolici (2019) emphasize, automation does not always entail replacing human effort entirely. Instead, it often involves diverse patterns such as decision support or blended decision-making, where human expertise remains integral. This perspective highlights a more nuanced view of AI's role, suggesting that its impact on skills depends on how systems are designed and integrated into tasks. At the same time, some applications demonstrate that AI's benefits are not uniform across all users. While certain tools primarily support novices, others prove more powerful for experienced users, not causing deskilling but rather enhancing their skills. Indeed, some applications might even need new skills to use effectively, another form of upskilling.

The question we seek to address in this paper is, under what conditions do these two outcomes emerge? What are the characteristics of tasks that when automated in particular ways lead to a levelling effect of technology versus those where technology better supports more experienced users? This question is important to identify the implications for workers as AI capabilities are built into more systems. The answer also has implications for how organizations might staff functions using the system and the longer-term implications of system usage.

Literature review

A common and long-standing predicted effect of computerization is deskilling, meaning the replacement of skilled workers by those with less skill or reduced opportunities for the same workers to exercise particular skills. Concern about deskilling has been raised since the dawn of computing (Mann & Williams, 1960; Whisler, 1970). Use of computer systems can strip a job of its content, leaving only a dull routine. For example, instead of solving a problem, a worker might instead feed relevant data to a computer and have it solve the problem. As a result, workers lose the opportunity or time to develop their skills through experimentation or on-the-job learning, or even to maintain skills previously acquired (Ardichvili, 2022; Li, Zhang, Niu, Chen, & Zhou, 2023). For instance, Rinta-Kahila et al. (2023) found that a company's reliance on an accounting package with sophisticated automation rendered its accountants—and consequently the organization as a whole—unable to perform a specific accounting process without the software, which they refer to as skill erosion. Organizational disruption ensued when the software was replaced with another, less automated system.

Deskilling has knock-on effects for the nature of the work, which can reinforce skill loss. As the flow of work becomes more like an assembly line, an individual worker's pace becomes regulated by the needs of processes on either side, and the need for interaction and resulting opportunity for social ties are reduced. Glenn and Feldberg (1977) described this process as the '*proletarianization of clerical work*'. They noted that even fifty years ago, clerical jobs were becoming more like factory jobs, with increased subdivision of work and specialization of workers due to automation and use of scientific management principles from classic organization theory

as management attempted to control workers and reduce the variability of their output. Zuboff (1988) pointed out that a system embodies assumptions about how the work should be done, resulting in a loss of flexibility for the worker. Formal rules replace discretion or specific knowledge, reducing workers' opportunities to display their mastery of their jobs. More recently, Holm and Lorenz (2022) found that when computers were used to give orders, the results for workers were increased work pace, constraints and decreased autonomy, an effect that was more pronounced for medium-skilled jobs. These changes in job content can lead to a loss of overview of the whole process (Ardichvili, 2022), further reducing workers' ability to learn and maintain appropriate skills.

The opposite prediction is upskilling. Computers can be used to automate the repetitive parts of a worker's job, leaving more interesting components for the human, and producing a more desirable job requiring a higher level of skills or having more responsibilities. For example, Zuboff (1988) presented a case in which the automation of a paper mill increased the role of the first-line production workers since they could control more than the single functions they used to. The jobs, therefore, required more skill, and the operators began to perform some of the functions of the managers. Even 60 years ago, Mann and Williams (1960) found some cases of job enlargement, noting that systems eliminated many routine jobs. Moreover, Sofia et al. (2023) (among many others) suggested that implementing AI will require new skills. They proposed that companies should help workers to identify which skills transfer and to develop needed new skills.

In practice, both effects, deskilling and upskilling, seem likely to occur simultaneously. There is some recent evidence from firm-level data of both effects. For example, Xue et al. (2022) found that Chinese companies reporting AI applications hire more employees without formal college education. However, McGuinness, Pouliakas, and Redmond (2023) found that skill-displacing technologies were positively associated with task variety and job-skill complexity, suggesting upskilling, though mostly for higher-skilled jobs. Zhang, Lai, and Gong (2024) also suggested an increase in employment for those with higher cognitive skills.

In past studies, deskilling or upskilling has often been viewed as dependent on deliberate choices about how to implement systems, driven by managers' preferences, e.g., for controlling versus working with workers (Zetka Jr, 1991). However, there is also a technical component to the decision as it is possible to design systems to promote hybrid intelligence (Rafner et al., 2022; Wahlström et al., 2024) and thus avoid deskilling. For example, Schemmer, Kühl, and Satzger (2022) described a decision support system that provides advice but requires users to make the final decision, thereby maintaining skill levels. Similarly, Arnold et al. (2023) designed a system with an interface based on expert knowledge representations and explanations, which improved novices' skills. And finally, the nature of the task itself is an important factor, interacting with managerial and technical impetuses.

Deskilling and upskilling due to AI

More recently, there have been a few studies that specifically include the differential impact of AI systems based on experience. We report on several that serve as the basis for our thinking. Table 1 summarizes these examples.

Brynjolfsson et al. (2023) reported on a study of an AI-based conversational assistant that supported the work of customer service agents by monitoring their chats with customers and suggesting possibly relevant documents to address the customers' problems. In a study with 5,179 customer support agents, they found that access to the tool boosted productivity, as indicated by a 14% increase in issues resolved per hour, while also increasing customer and worker satisfaction. However, the productivity increase was restricted to novice and low-skilled workers, who saw a 34% improvement; experienced and highly skilled workers experienced minimal benefit. They

Paper	Input	Evaluation	Editing	Impact
Brynjolfsson et al. (2023)	Extracted from customer chat	Relevance of document to problem	None	Deskilling
Dell'Acqua et al. (2023)	Prompt for problem to solve	Evaluation of suitability of suggestions	Output text lightly edited	Deskilling
Noy and Zhang (2023)	Used task prompts unchanged	Evaluate suitability of output	Output text lightly or not edited	Deskilling
Campero et al. (2022)	Prompt for UI element	Visual evaluation of appearance	Graphical interface to position; option to edit code	Deskilling
Peng et al. (2023) and Cui et al. (2024)	Prompt for needed code	Evaluate suitability of code	Edit code to fix bugs and adapt to need	Deskilling [a]
Luo et al. (2021)	Extracted from calls	Evaluate coaching advice	Implement coaching advice	Deskilling [b]
Wang et al. (2023)	Extracted from medical record	Determine suitability of billing code	Accept or reject code and possibly add others	Skill maintaining
Choudhury, Starr, and Agarwal (2020)	Extracted from patent application	Determine suitability of search terms	Possibly add additional search terms	Skill maintaining
Dell'Acqua (2024)	Extracted from the job application	Evaluate system suggestion	Accept or reject suggestion	Skill maintaining [c]
Kim and Kang (2024)	Preset inputs	Evaluate recommendation and important factors	Incorporate factors in report	Skill maintaining
Jia, Luo, Fang, and Liao (2024)	Preset inputs	None	None	Automation & upskilling

[a] Not statistically significant
[b] When volume of suggestions matched to user abilities
[c] With less accurate system

Table 1. Summary of results from the literature.

suggested that the AI model spreads the best practices of more proficient workers, which we interpret as evidence for deskilling because a worker need not be as skilled to perform well.

Dell'Acqua et al. (2023) reported on two experiments with Boston Consulting Group consultants. We focus on the first, in which 385 consultants carried out a set of 18 realistic consulting tasks designed to be within the capabilities of AI, namely, to conceptualize and develop new product ideas. Consultants either had no AI support, access to ChatGPT-4 or access to ChatGPT-4 with a prompt engineering overview. Consultants with access to ChatGPT were more productive and produced higher quality output, with a much stronger effect for consultants who performed lower on an initial assessment task. Those who received training in prompt engineering performed somewhat better than those who did not. These results show levelling and deskilling, as the system enables those who previously displayed less skill to operate at higher level.

In an experimental study with a writing task carried out by 453 college-educated professionals, Noy and Zhang (2023) found that those using ChatGPT both saved time and increased quality. Subjects with access to ChatGPT were given examples of prompts as a form of training. The authors reported that subjects seem to have mostly used ChatGPT's output as is, with little or no editing. Those who scored worse on an initial task improved their quality more, again evidence for levelling and deskilling.

Campero et al. (2022) explored having 200 programmers develop HTML code to replicate a web page, half using ChatGPT-3 with prior conditioning to generate relevant HTML code. The users had a graphical interface with which they could reposition the element and could edit the

generated HTML if desired. The programmers using ChatGPT completed the task about 30% faster. Interestingly, when they had 50 non-programmers do the same task with ChatGPT, they found that 95% of them finished the task in about the same time as the professional programmers. They concluded that this use of AI can '*be seen as a form of deskilling for the programmers whose jobs could now be performed by people with less skill—and for lower compensation*'.

Peng et al. (2023) reported on a study of the productivity impacts of GitHub Copilot in which 95 programmers were recruited to write a simple HTTP server in JavaScript, 45 using Copilot and 50 without it. The treatment group finished in less than half the time, with roughly the same level of success. Though the effect was only marginally significant, they also found that developers with fewer years of experience benefited more, evidence again for deskilling.

Cui et al. (2024) reported on three experiments conducted by three large companies who randomly assigned developers to use CoPilot. They concluded that use of the tool leads to a 26% increase in weekly coding tasks completed over all three companies. Focusing just on Microsoft, they found that CoPilot adoption was higher for junior developers and developers with lower tenure in the company. Further, the increase in pull requests, commits and builds was roughly three times higher for lower tenure developers, though the differences were not statistically significant. Overall, these results seem consistent with deskilling.

Luo et al. (2021) reported on an AI coaching system for sales representatives. The system analyses the agents' calls to give advice about improving the interaction with customers. From an experiment with 429 agents, it was found that the system helped middle-ranked agents increase their sales rate the most, to nearly the level of higher-ranked agents. Lower-ranked agents were unable to absorb the volume of suggestions, while higher-ranked agents were averse to AI-generated advice. When the volume of suggestions was reduced, lower-ranked agents also improved, i.e., further levelling.

Wang, Gao, and Agarwal (2023) reported on the effects of an AI system to support coding of medical records. From a study with 80 coders using the system and 468 in the control group, they found that the system increased the productivity of all workers (reportedly with no impact on quality), but more so for those with more task experience, who could more quickly evaluate the proposed codes for suitability. They noted that '*If an AI is successfully trained on a task-specific data set, AI can substitute for a worker's task experience*'. However, in their study, the benefit went to those with more task experience. On the other hand, the system does not seem to have changed the nature of the skills required. In summary, this study seems to find neither deskilling nor upskilling, rather maintaining the advantage of having more experience.

Choudhury et al. (2020) examined how expert knowledge can address system shortcomings. Specifically, they studied a system that processes patent applications to suggest search terms to find relevant existing patents. Since patent applications may be deliberately designed to look different from the prior art, an experienced human patent examiner can complement the system by expanding the search. In an experiment, 221 MBA students examined five patent claims that were invalidate by an existing patent that used different language than the application. To simulate expertise, half of the subjects were given expert advice about how to search that included advice on adding terms. The experiment showed that those using the new system found a more precise set of relevant patents that were more like the application, as intended, but that the search was unlikely to find the relevant patent. The expert advice made it more likely that the patent would be found, again suggesting the importance of the existing expertise for this task.

Kim and Kang (2024) studied 97 mutual fund analysts who write reports rating mutual funds including an explanation for the rating. Half had access to a proprietary rating algorithm that rated the fund and identified the factors that were important for the prediction. They found that access to the predictions improved recommendation quality (i.e., whether the prediction matched the

outcome) for simple cases, but had a negative impact on explanation quality, especially for junior analysts. New analysts were less likely to have algorithmic aversion but found it hard to incorporate the system results into their thinking and so wrote shorter reports that were less coherent and included more causal drivers in the explanation. We interpret this case as the system better supporting more skilled users, since skill was needed to make proper use of the system outputs.

In an experiment with recruiters evaluating job applications, Dell'Acqua (2024) varied the quality of the AI support provided. They found that expert recruiters using a less accurate system were more likely to carefully evaluate the applications themselves rather than simply taking the system suggestion, resulting in a more accurate evaluation than those using the more accurate (but imperfect) system without a careful evaluation. Indeed, more experienced recruiters using the better system performed worse than less experienced recruiters. They conclude that '*an AI that is 'too good' may induce workers to mindlessly follow algorithmic advice and lead to over-delegation*' and suggest that '*collaboration should be designed with the goal of keeping humans attentive in tasks where their focus is necessary to improve performance*'.

Finally, Jia et al. (2024) studied 40 sales agents interacting with 3144 potential customers to sell credit cards, working in two phases: first qualifying leads by assessing interest and then engaging to make a sale. Half of the agents used an AI telephone conversational system that autonomously did the first step, while the other half did it themselves. They found that agents using the AI system were more likely to make a sale because the system screened out likely-uninterested leads, allowing them to focus on better prospects. However, top agents were 2.8 times more likely to make a sale than bottom agents, which they attributed to the top agents' ability to develop better sales scripts and to answer questions for which they had not been trained, which bottom agents did not do. This case is evidence for upskilling: by taking over the routine part of a job, the system leaves work that requires more skill to perform at a high-level.

Model development

As a basis for analysing different applications of AI, we propose a simple model of the interaction among human, technology and task. In our model, the user performs a task that involves problem assessment and the creation of some output. For the scope of this paper, we focus on information tasks, not physical tasks, covering a broad category such as decision-making, customer care, brainstorming of ideas, etc. The model structure aligns with Crowston and Bolici's (2020) framework, which identifies three patterns of machine learning use—decision support, blended decision-making, and complete automation—and highlights how automation can affect not only specific tasks but also interdependent processes and coordination mechanisms. When using a system to support a task, rather than performing the task directly, users follow a process including: 1) assessing the task that should be executed, 2) possibly formulating an input and providing it to the AI system, 3) assessing the result, 4) accepting, regenerating, or editing the output, and 5) completing the task. For example, a human interacting with a document repository to find an answer to a problem will formulate a query (or use a query generated by the system), look at results to see if they meet the requirements, pick one or redo the query and try again. For interaction with a large language model (LLM) such as ChatGPT, the human will formulate a prompt, evaluate the generated results, tweak the prompt if the results are unsatisfactory, and possibly edit the output to improve it to complete the task.

Model components

Understanding the deskilling or upskilling impacts of AI requires a comprehensive model that captures the interaction between four main elements: 1) Humans, 2) Systems, 3) the Outputs generated, and 4) the Tasks that must be performed in the organization by humans and/or system. It is based on the previous research on the roles of expertise, prompting, system accuracy and task nature in assessing AI-enabled work. For example, Zuboff (1988) observed that systems reflect work

assumptions, limiting flexibility and eroding skills (Human–Task; Human–AI), while Rinta-Kahila et al. (2023) demonstrated skill erosion from over-reliance on automation (Human–AI; Human–Task). Conversely, application of hybrid intelligence (Schemmer, Kühl and Satzger, 2022) mitigates deskilling by involving humans in decision-making (Human–AI; Human–Task). Brynjolfsson et al. (2023) illustrated a leveling effect, enabling novices to reach intermediate performance, though not expert levels (Human–Outputs; Human–Task). Wang, Gao, and Agarwal (2023) emphasized human expertise in refining AI outputs, underscoring the dual-edged nature of AI's impact on skills and performance (Human–Outputs; Human–AI). Our model is intended to analyse these interactions and their implications for skill development, use and retention. The proposed model highlights the interplay between human expertise, system capabilities, and task requirements in shaping task performance outcomes.

1) *HUMAN*: the persons that must perform the task and who are deciding if and how to use a system to support or to substitute for their work. We focus on two main characteristics that have an impact on the process:
   a) Domain knowledge: the extent to which the human is an expert in the specific domain relevant to the task. Higher domain knowledge enables better assessment and refinement of system outputs.
   b) Input formulation knowledge: the human's ability to effectively formulate inputs for the AI system. For instance, when using an LLM, expertise in prompting can significantly influence the quality and relevance of the system's outputs.

2) *AI*: the specific system that can be accessed by the human during the task execution and for which we consider three characteristics:
   a) Input variability: The variability of input to the system. Some systems take a fixed set of variables while in contrast, an LLM can take nearly any text as input.
   b) Accuracy/limitations: The constraints of the system, such as the propensity for generating errors or the need for human intervention to correct and refine outputs.

3) *OUTPUTS* of the system: the answers that the system provides in response to the input that have at least the following characteristics:
   a) Quality: The accuracy, relevance, and usability of the system-generated outputs. High-quality outputs require less modification and are more useful for completing the task.
   b) Speed: We presume that the system will be able to generate an answer more quickly than the human, leading to the observed increases in speed.

4) *TASK*: the activity that must be performed and for which we can consider:
   a) Nature of the task: the specific characteristics of the task, including whether it is creative, analytical, or procedural.
   b) Task division: how the task is split between the human and the system. This could involve the human performing the entire task, the system performing the entire task, or a collaborative effort where both the human and system contribute.

Model phases

Considering the model as a representation of the dynamics of interaction between the various elements of the process, we can distinguish four different temporal phases.

1) *Phase One*: the human responsible for carrying out the task can decide whether to be supported by a system and, if so, in what way. It may be though that usage is non-discretionary, meaning the human user is obligated to use the system.

2) *Phase Two*: the human utilizes their domain knowledge and skills to interact with the system to obtain support. It may be that the inputs to the system are predetermined by the task, or the user may have freedom to craft an input.

3) *Phase Three*: the results generated by the system are assessed and interpreted by the human, who decides whether to accept them or refine either manually or through further through additional interaction with the system.

4) *Phase Four*: the final results are used to execute the task, either in support of or as a substitute for direct human involvement.

Interaction among the components

The interplay between the components of the model determines the overall impact on skill levels and task performance.

- *Human and System Interaction*: the effectiveness of the system may be dependent on the human's ability to craft an input. Those with high knowledge can generate better initial responses from the system or be better in refining the outputs iteratively tuning the prompting itself. Indeed, it may be that the prompts are created by experts who develop a system rather than by the end-user using the system for a task.
- *System and Output*: the system's capabilities and limitations directly affect the quality and adaptability of the outputs. High-quality outputs reduce the need for extensive human intervention and can be applied to a variety of tasks, enhancing productivity.
- *Human and Output*: the human's role in assessing and interpreting the system's output depends (again) on their domain knowledge. High domain knowledge allows for quicker and more accurate assessment of the output, reducing the risk of simply accepting a wrong or incomplete result. Experts in the domain can better refine system outputs if they are not suitable.
- *Outputs and Task*: the nature of the outputs influences how the task is performed. High-quality, adaptable outputs can enhance productivity and potentially upskill workers by allowing them to focus on higher-level refinements. On the other hand, poor outputs can lead to deskilling if the human's role is reduced to merely accepting or rejecting system-generated content without substantial engagement.

This phased approach highlights the iterative and interactive nature of the model, emphasizing the crucial role of human expertise at each stage to maximize the effectiveness of the LLM and ensure the successful completion of the task. The conclusion of the model can lead to different impacts on the need for expertise.

- *No Effect*: in this scenario, the use of AI has no impact on the skills of the individuals involved. The task is performed similarly whether or not the AI is used, and the human's existing knowledge and skills remain unchanged. However, the system may have other benefits, e.g., for speed or quality.
- *Levelling Effect*: this scenario occurs when AI minimizes the importance of the human's knowledge on task performance. The use of AI flattens the importance of prior knowledge, as a novice using AI can achieve a task performance like that of an expert. In this case, the AI levels the playing field, reducing the skill gap between novices and experts.
- *Multiplier Effect*: in this scenario, the use of AI acts as a multiplier on the human's existing knowledge, thereby increasing the performance gap between novices and experts. The AI enhances the capabilities of those with higher prior knowledge, leading to significantly better task performance compared to novices. This effect underscores the role of AI in amplifying the skills and expertise of experienced users.

By understanding these different scenarios, we can better anticipate the implications of AI integration into various workflows and design strategies to optimize both human and AI contributions to task performance.

We speculate that a system with pre-formed prompt with results that are easy to assess and that have little need to edit more likely results in levelling and so deskilling, as a more expert worker does not have an opportunity to employ their expertise. On the other hand, a system could have more flexibility about prompting or more need for output assessment and editing, tasks that experts could be quicker and more accurate in doing. To the extent that the task has these characteristics, it is more likely to benefit from expertise and thus help experts more than non-experts.

Findings: deskilling and upskilling due to AI

We illustrate our model by analysing some of the studies surveyed above. For instance, in Brynjolfsson et al. (2023)'s study, the prompt is taken from the customer chat, not the agent. The agent needs to assess if a proposed document is apropos but can also suggest it and let the customer assess. If appropriate, the solution is provided to customer as is. Therefore, our model suggests that the effect of the system will be levelling, as found: the system can provide solutions that a more experienced employee would suggest, but without requiring the same level of expertise. Similarly, in Noy and Zhang (2023)'s study, subjects using ChatGPT seem to have copied the writing prompts from the problem and used ChatGPT's output largely unchanged. They had to evaluate if the output was suitable but given the similarity of the task to their regular work, we expect this evaluation to be straightforward (that is, subjects differed in the quality of their writing, but we think not in the ability to assess suitability of output).

In about half of our examples, the input to the system was generated from the task, eliminating the need or ability of a user to develop expertise in directing the system. However, these cases include our two examples of skill maintenance. In other cases, users reportedly used the given task prompts unchanged, again not developing skills. In two of the cases we reviewed, users could prompt more freely (Dell'Acqua et al., 2023; Peng et al., 2023). Interestingly though, those cases still resulted in deskilling. In summary, the cases reviewed do not suggest that the ability to better prompt an AI system is as yet a distinguishing characteristic in task performance.

In contrast, ability to evaluate and make use of a system's output does seem to play a role in several cases. Wang et al. (2023)'s study poses an interesting example. In this case, the search is based on sentences in the medical record. However, the authors report that evaluation of the suggestions was required to rule out false positives, which was quicker for more experienced workers. In this case, the system does not require new skills (e.g., for prompting) and maintains the value of existing skills (evaluation). In Kim and Kang (2024)'s study, expertise was seemingly needed to evaluate the system's output and to incorporate it into the final product, making the system more useful for more experienced workers. Choudhury et al. (2020) found that the search terms identified by the system needed to be augmented, which required expertise to do successfully. Luo et al. (2021) found an inverted U-shape effect for the initial coaching system: it did not help experienced workers, who already knew the job, nor inexperienced workers who could not cope with the volume of suggestions, but did help people in between, the later effect again highlighting the importance of being able to assess and incorporate system output.

Finally, the case of Jia et al. (2024) is one of few studies we found that reportedly resulted in upskilling. Interestingly, this case is also one in which one subtask was completely automated, namely the initial screening call with a potential customer, while the remaining subtask left to the human is performed without support, a subtask for which greater skill translates into better performance. (In other words, the analysis in Table 1 does not describe the task the human

performs.) We are curious what the impact would be of supporting the sales task and speculate that it could lead to levelling, as found by Luo et al. (2021).

Overall, we perceive a general pattern: if you have too little skill, you can't make use of the system outputs. If you have moderate skill, the system generally seems to help achieve better performance. If you have a lot of skill, the system doesn't help as much and may even be resisted (Wang et al., 2023).

Discussion

The model and the studies reviewed more broadly suggest several points for consideration.

First, reviewing the papers identified, we note a lacuna, namely we identified no studies in which better prompting skills gave more experienced workers an advantage. We expected that using Copilot to support programming would have these effects. For instance, Mozannar et al. (2024) observed that programmers using CoPilot spent over 20% of their time thinking about or verifying a CoPilot suggestion, about 10% of the time editing a suggestion, and about 10% crafting prompts. Prompt crafting is often iterative: write a prompt, assess output, tweak the prompt. Often, suggestions were accepted to fully evaluate and tweak them, not necessarily because they were correct. Dibia et al. (2022) found that experienced programmers still found incorrect code suggestions from CoPilot useful even if the code was not entirely correct, as it could be modified with little effort, thereby increasing productivity. Similarly, Zamfirescu-Pereira et al. (2023) found that while the code generated by CoPilot often had errors, they were easier to fix than errors in code generated by humans. They concluded that '*CoPilot can become a liability if it is used by novice developers who may fail to filter its buggy or non-optimal solutions due to a lack of expertise*' (Zamfirescu-Pereira et al., 2023). Randazzo et al. (2024) suggests that ChatGPT users who retain overall control of the task, strategically deciding which tasks to delegate, perform better than those who direct the system through the whole task and much better than those who delegate entirely. However, we only found two studies (Cui et al., 2024; Peng et al., 2023) that examined the impacts of individual differences using this technology and unfortunately, these studies do not provide much detail about how developers interacted with the system.

Second, questions about deskilling and upskilling have important organizational implications that need to be considered. For instance, if the system results in deskilling, organizations may be tempted to hire less skilled workers or to invest less in training since performance with the system will still be satisfactory. These temptations will likely be greater for jobs that face high turnover, such as customer support. A consideration is that managers tend to systematically underestimate the expertise needed to do the work of their employees, meaning that they may classify more work as replaceable or low skilled than is appropriate. This consideration reinforces the importance of involving the people doing the work in system design. A further consideration is the implications for organizational learning. If the problem is not static, but the system has a levelling effect, then who will learn the answers to the new questions, if there are no longer any experts doing the tasks? Relatedly, it is an open question whether non-experts using a system learn to do the task independently or whether the system obviates the need to learn.

On the flip side, systems that reward expertise also raise concerns. If expertise is more valued, organizations need to consider how it is developed. For instance, there are anecdotal reports of companies no longer hiring entry-level workers to do what LLMs can do (Edwards, 2023; Yegge, 2024). If the work of entry-level positions can be largely automated, organizations will face the problem of how new hires develop the necessary expertise to oversee the automated work.

Third, system use may require user to develop new skills in prompt crafting and in evaluation and use of system output, rather than manual creation of output. There is some evidence for these effects, e.g., the small improvements found by Dell'Acqua et al. (2023) for the short prompt-crafting

training, and the several studies in which expertise was needed to evaluate outputs (Choudhury et al., 2020; Kim & Kang, 2024; Luo et al., 2021; Wang et al., 2023). However, Dell'Acqua (2024) raise the issue of needing to motivate workers to be critical about systems, more critical ironically for systems that work better.

Fourth, the future of work with AI and the related necessary skills requires consideration of the inherent nature of AI, which is best able to provide answers for problems and solutions that frequently appear in its training data. This limitation could lead to a need for workers who have skills primarily in identifying corner cases and their possible solutions that the AI cannot handle. Understanding and designing how to support these types of skills in workers remains an open question. If the management (identification of problems and solutions) of the most common cases is done through or with AI, it cannot represent a learning field for new workers who will need to learn to handle specific and less frequent corner cases.

Fifth, our model also posits limits to the impacts of technology support. Specifically, Amdahl's law (Amdahl, 1967) says that speed up due to a new system component is limited by fraction of time the using new component. As an example, if only 10% of a job is automated, the maximum speed-up is 1/90% or about an 11% speed-up. Reasoning in reverse, to make someone 2x faster at their work (i.e., a multiplier effect), as found by Peng et al. (2023) for the programmers using ChatGPT, requires eliminating 50% of what they do. We speculate that such a result implies that programmers are seeing benefits by having the system write entire functions at a time, rather than writing lines of code. Our model does not as yet capture the possibly transformative effects of entirely changing the nature of the task performed.

Finally, our model has at least one design implication. As prompting is a new skill that has a potential to make a difference to the results, it might be beneficial to let users tweak the prompts, even if they are mostly preset or derived from the task data. The visibility may help people to develop expertise in prompting and use this new skill to improve results.

Conclusion

We conclude with some ideas for future research. First, this model is based on examination of a few sample implementations of AI to support work. More systematic studies across a wider range of tasks would help refine it and demonstrate its utility. To carry out these studies will require more details about the nature of the task, the technology and the workers' interactions. It would be helpful to have more detail about the specifics of the systems and how people interact. We would like to dig into the details of the system more to understand where skills make a difference. For instance, it could be that crafting a good query for a search is a more important skill than getting an LLM prompt exactly right, given the latter's flexibility and interpretive abilities.

Second, use of the model could guide studies of new systems. For instance, it would be interesting to vary the level of prompt crafting possible for the same task and exploring impact on workers with varying skill levels. We expect more-skilled workers to use the capabilities to further extend their advantage over less-skilled workers, but this is an empirical question that needs study.

Third, it is important to consider that this model, like most studies focused on AI to date, focused on the relationship between work and AI within a single task. However, work is a process made up of multiple interdependent tasks. Therefore, the potential impact of AI on work should be examined within the context of how AI is used across a set of tasks that need to be coordinated. This broader perspective acknowledges that the integration of AI affects not only individual tasks but also the overall workflow, requiring a broader approach to understand its full implications on job performance and skill requirements.

Acknowledgements

Kevin Crowston was partly supported by a grant from the United States National Science Foundation, Grant 21-29047.

About the authors

Kevin Crowston is a Professor of Information Science in the School of Information Studies, Syracuse University, Syracuse, New York, USA. He received his Ph.D. from the Massachusetts Institute of Technology and his research interests are in the futures of work using advanced technologies. He can be contacted at crowston@syr.edu

Francesco Bolici is a Professor of Organization Studies and Information Systems and Director of OrgLab at the Department of Economics and Law, Università degli studi di Cassino, Italy. Holding a Ph.D. from Luiss Guido Carli University, Rome, his research examines how information technology enables new possibilities for reimagining work design and organizational practices. He can be contacted at f.bolici@unicas.it

References

Amdahl, G. M. (1967, April 18-20). Validity of the single processor approach to achieving large scale computing capabilities. In Proceedings of the Spring Joint Computer Conference (pp. 483–485).

Ardichvili, A. (2022). The impact of artificial intelligence on expertise development: Implications for HRD. Advances in Developing Human Resources, 24 (2), 78–98. https://dx.doi.org/10.1177/15234223221077304

Arnold, V., Collier, P. A., Leech, S. A., Rose, J. M., & Sutton, S. G. (2023). Can knowledge-based systems be designed to counteract deskilling effects? International Journal of Accounting Information Systems, 50, 100638. https://dx.doi.org/10.1016/j.accinf.2023.100638

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative ai at work (Tech. Rep. No. Working Paper 31161). National Bureau of Economic Research. Campero, A., Vaccaro, M., Song, J., Wen, H., Almaatouq, A., & Malone, T. W. (2022). A test for evaluating performance in human-computer systems (https://arxiv.org/abs/2206.12390).

Choudhury, P., Starr, E., & Agarwal, R. (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation. Strategic Management Journal, 41 (8), 1381-1411. https://dx.doi.org/10.1002/smj.3152

Crowston, K. & Bolici, F. (2019) Impacts of machine learning on work. In Proceedings of the Hawaii International Conference on System Sciences, Hawai'i, USA. https://dx.doi.org/10.24251/HICSS.2019.719

Crowston, K., & Bolici, F. (2020). Impacts of the Use of Machine Learning on Work Design. In Proceedings of the 8th International Conference on Human-Agent Interaction (pp. 163-170). https://doi.org/10.1145/3406499.3415070

Cui, Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2024). The effects of generative ai on high skilled work: Evidence from three field experiments with software developers. Social Science Research Network (SSRN). https://dx.doi.org/10.2139/ssrn.4945566

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Tech. Rep. No. 24-013). Harvard Business School Technology & Operations Management Unit. (https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf)

Dell'Acqua, F. (2024). Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Laboratory for Innovation Science, Harvard Business School. (https://www.fabriziodellacqua.com/)

Dibia, V., Fourney, A., Bansal, G., Poursabzi-Sangdeh, F., Liu, H., & Amershi, S. (2022). Aligning offline metrics and human judgments of value of AI-pair programmers (https://arxiv.org/abs/2210.16494).

Edwards, B. (2023). IBM plans to replace 7,800 jobs with AI over time, pauses hiring certain positions. (https://arstechnica.com/information-technology/2023/05/ibm-pauses-hiring-around-7800-ro)

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. Information and Organization, 28 (1), 62-70. https://dx.doi.org/10.1016/j.infoandorg.2018.02.005

Glenn, E. N., & Feldberg, R. L. (1977). Degraded and deskilled: The proletarianization of clerical work. Social Problems, 25 (1), 52–64.

Holm, J. R., & Lorenz, E. (2022). The impact of artificial intelligence on skills at work in Denmark. New Technology, Work and Employment, 37 (1), 79–101.

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. Academy of Management Journal, 67 (1), 5–32.

Kim, H., & Kang, X. (2024, Aug). Machine predictions and causal explanations: Evidence from a field experiment. Symposium presentation at the Academy of Management Meeting, Chicago, IL, USA. https://dx.doi.org/10.5465/AMPROC.2024.13661symposium

Li, C., Zhang, Y., Niu, X., Chen, F., & Zhou, H. (2023). Does artificial intelligence promote or inhibit on-the-job learning? Human reactions to AI at work. Systems, 11 (3), 114.

Luo, X., Qin, M. S., Fang, Z., & Qu, Z. (2021). Artificial intelligence coaches for sales agents: Caveats and solutions. Journal of Marketing, 85 (2), 14-32. https://dx.doi.org/10.1177/0022242920956676

Mann, F. C., & Williams, L. K. (1960). Observations on the dynamics of a change to electronic data-processing equipment. Administrative Science Quarterly, 217–256.

McGuinness, S., Pouliakas, K., & Redmond, P. (2023). Skills-displacing technological change and its impact on jobs: Challenging technological alarmism? Economics of Innovation and New Technology, 32 (3), 370–392.

Mozannar, H., Bansal, G., Fourney, A., & Horvitz, E. (2024). Reading between the lines: Modelling user behaviour and costs in AI-assisted programming. In Proceedings of the CHI Conference on Human Factors in Computing Systems. https://dx.doi.org/10.1145/3613904.3641936

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381 (6654), 187–192.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot (https://arxiv.org/abs/2302.06590).

Rafner, J., Dellermann, D., Hjorth, A., Veraszto, D., Kampf, C., MacKay, W., & Sherson, J. (2022). Deskilling, upskilling, and reskilling: A case for hybrid intelligence. Morals & Machines, 1 (2), 24–39.

Randazzo, S., Lifshitz-Assaf, H., Kellogg, K., Dell'Acqua, F., Mollick, E. R., & Lakhani, K. R. (2024). Cyborgs, centaurs and self automators: Human-genAI fused, directed and abdicated knowledge co-creation processes and their implications for skilling. (https://ssrn.com/abstract=4921696)

Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W., & Ruissalo, J. (2023). The vicious circles of skill erosion: A case study of cognitive automation. Journal of the Association for Information Systems, 24 (5), 1378–1412.

Schemmer, M., Kühl, N., & Satzger, G. (2022). Intelligent decision assistance versus automated decision-making: Enhancing knowledge workers through explainable artificial intelligence. In Proceedings of the 55th Hawai'i International Conference on System Sciences. Virtual. (https://hdl.handle.net/10125/79517)

Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT: More than a "weapon of mass deception": Ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. International Journal of Human–Computer Interaction, 1–20.

Sofia, M., Fraboni, F., De Angelis, M., Puzzo, G., Giusino, D., Pietrantoni, L. et al. (2023). The impact of artificial intelligence on workers' skills: Upskilling and reskilling in organisations. Informing Science: The International Journal of an Emerging Trans-discipline, 26, 39–68.

Wahlström, M., Tammentie, B., Salonen, T.-T., & Karvonen, A. (2024). AI and the transformation of industrial work: Hybrid intelligence vs double-black box effect. Applied Ergonomics, 118, 104271.

Wang, W., Gao, G., & Agarwal, R. (In press). Friend or foe? Teaming between artificial intelligence and workers with variation in experience. Management Science. https://dx.doi.org/10.1287/mnsc.2021.00588

Whisler, T. L. (1970). Information Technology and Organizational Change. Belmont, CA: Wadsworth.

Xue, M., Cao, X., Feng, X., Gu, B., & Zhang, Y. (2022). Is college education less necessary with AI? Evidence from firm-level labor structure changes. Journal of Management Information Systems, 39 (3), 865–905.

Yegge, S. (2024). The death of the junior developer. (https://sourcegraph.com/blog/the-death-of-the-junior-developer)

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. https://dx.doi.org/10.1145/3544548.3581388

Zetka Jr, J. R. (1991). Automated technologies, institutional environments, and skilled labor processes: Toward an institutional theory of automation outcomes. Sociological Quarterly, 32 (4), 557–574.

Zhang, W., Lai, K.-H., & Gong, Q. (2024). The future of the labor force: Higher cognition and more skills. Humanities and Social Sciences Communications, 11 (1), 1–9.

Zuboff, S. (1988). In the Age of the Smart Machine: The Future of Work and Power. Basic Books, Inc.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Inconsistency-driven approach for human-in-the-loop entity matching

Hiroyoshi Ito, Takahiro Koizumi, Ryuji Yoshimoto, Yukihiro Fukushima, Takashi Harada, and Atsuyuki Morishima

DOI: https://doi.org/10.47989/ir30iConf47140

Abstract

Introduction. Entity matching is a fundamental operation in a wide range of information management applications and a tremendous number of methods have been proposed to address the problem. Human-in-the-loop entity matching is a human-AI collaborative approach which is effective when the data for entity matching is incomplete or requires domain knowledge. A typical human-in-the-loop approach is to allow a machine-learning-based matcher to ask humans to match entities when it cannot match them with high confidence. However, ML-based matchers cannot avoid the unknown-unknown problem, i.e., they can resolve the entities incorrectly with high confidence.

Method. This paper addresses an inconsistency-based method to deal with this problem. The method asks humans to resolve the entities when we find inconsistency in the transitivity property behind entity matching. For example, if a matcher returns a positive result only for two combinations among three entities, the result is inconsistent.

Analysis. This paper shows an implementation of our idea in similarity-based blocking method and Bayesian inference and explains the result of an extensive set of experiments that reveals how and when the method is effective.

Results. The result showed that the inconsistency-based sampling selects very different entity pairs compared to other sampling strategies and that a simple hybrid strategy performs well in many practical situations.

Conclusion. The results indicate our approach complements any existing matcher that can cause the unknown-unknown problem in entity matching.

Introduction

Entity matching is a fundamental operation for objects in information management applications, such as bibliographic records, names (Cohen, et al., 2003), entities that appear in ontology (Xu, et al., 2008), texts and other data collections (Jaro, 1989). Therefore, a tremendous number of methods have been proposed to address the problem (Christophides, et al., 2020), (Mudgal, et al., 2018), (Jaro, 1989), and implemented services are often available (Govind, 2018). Machine-learning-based matchers (Eraheem, et al., 2014), (Li, et al., 2020), (Yao, et al., 2022) are widely used for many other problems in digital libraries (Nielsen, 2018).

Human-in-the-loop entity matching (Osawa, et al., 2021), (Gokhale, et al., 2014), (Das, et al., 2017) is a human-AI collaborative approach to the problem and known as being effective when the data set is incomplete, or the matching requires domain knowledge (Trabelsi, et al., 2022). A typical human-in-the-loop ML matching is to allow an ML matcher to ask humans to match entities when it cannot match them with high confidence.

However, ML-based matchers cannot avoid the *unknown-unknown problem*, i.e., they can match the entities incorrectly with high confidence (Chung, et al., 2019). This is an inherent weakness of the ML-based matchers, because typical human-in-the-loop approaches choose a pair when the matching result is uncertain, or multiple matchers disagree on the result (Settles, 2010), but such procedures do not guarantee that the matching decisions on the remaining pairs are correct.

This paper addresses an inconsistency-driven method that addresses this problem, which can be used with *any* matchers that output matching probabilities for a pair of entities. The method chooses the entity pairs it asks humans to match, by an inconsistency-based "sampling"; when it finds inconsistency in the equivalence relation behind entity matching, it picks up the pair of entities that cause the inconsistency and asks humans to fix the matches. For example, if a matcher returns a negative result only for a particular pair (e.g., s_1 and s_3) among three entities s_1, s_2 and s_3 (Fig 1), the result is inconsistent because it violates the transitive law of the equivalence relationship.

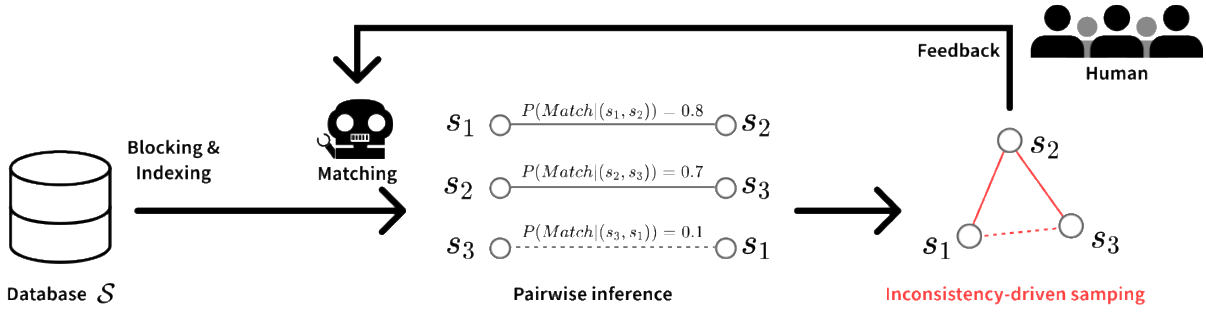


Figure 1. Inconsistency-driven human-in-the-loop entity matching: if the ML matcher outputs the results that are inconsistent with each other, humans correct the result, and the feedback is given to the matcher to improve the results

We implemented the inconsistency-driven sampling in a simple framework combining Bayesian inference and similarity-based blocking method, to highlight the effects of the sampling. In the framework, we assume that we have a certain amount of training data with known labels in advance for the blocking and Bayesian inference. Then, we conducted an extensive set of experiments with different accuracies of matchers, expecting that the result reveals how and when the method is effective. Therefore, we focus on the quality improvement relative to the performance (in terms of f1 values (We use the f1 value instead of accuracy because of the imbalance of the numbers between matched and unmatched pairs)) of the original matcher.

Our research questions are as follows: (RQ1) What is the characteristic of the inconsistency-driven approach as an entity matching strategy? (RQ2) In what situation the approach is effective?

The contributions of this paper are twofold: First, we show a principled framework for an inconsistency-driven approach for human-in-the-loop entity matching. Our approach complements any existing matcher that can cause the unknown-unknown problem. This paper shows an implementation of our approach in a simple framework to address the effects of the approach.

Second, we show the result of an extensive set of experiments with three real-world datasets with different characteristics. We then conducted a detailed analysis of the results. Consequently, we observed the effects of our approach in different situations, revealing how and when the approach is effective.

Note that our research question is not about the performance of a particular matcher. Our findings are summarized as follows.

1. The inconsistency-driven sampling chooses data items that are completely different from those chosen by uncertainty-based and random sampling.
2. The inconsistency-driven sampling has advantages even if human answers are not completely correct as long as the performance of the matcher is relatively high. This is because the inconsistency-based check works well in finding incorrect responses from humans as long as the performance of the matcher is relatively high.
3. A simple hybrid strategy broadens the sweet spot of the inconsistency-driven approach; it lowers the f1 value threshold for which inconsistency-driven sampling is effective.

Related works

In this section, we introduce the studies related to this research and describe their methods and the position of this research.

Rule-based or ML-based entity matching

Many solutions have been proposed for the entity matching problem. The classic example is a rule-based approach. The method clusters those that contain full or partial matches of attribute values or identical tokens that have been segmented into words (Jaro, 1989), (Cohen, et al., 2003), (Benjelloun, et al., 2009). Recently, a machine learning approach has been proposed. They use random forests (Gokhale, et al., 2014), (Das, et al., 2017), and metric learning (Peeters, et al., 2022), (Osawa, et al., 2021) to judge whether an entity pair is a match by calculating its similarity. These methods are simple and inexpensive but are vulnerable to orthographical variants of data.

Human-in-the-loop entity matching

Several studies have pointed out the limitations of entity matching using only computers without human intervention. (Takashi et al. 2019) identified low-precision results based on similarity using Okapi BM25 (Robertson, et al., 1995) and proposed a crowdsourcing-based method that allows for human interaction. (Das, et al., 2017) proposed the framework, Falcon. This method assumes that there is no training data, and a small portion of the target data is labelled by a human to prepare the training data. Other methods using human-in-the-loop have been proposed (Li, 2017), (Gokhale, et al., 2014), (Osawa, et al., 2021), (Eraheem, et al., 2014). Any methods are used to generate or modify data for quality of results and training of the computer, and the other methods are used in much the same way. In this study, we propose a human task from the perspective of obtaining efficient training data.

Methods using transitivity law and domain knowledge

(Zhu et al., 2020) proposed a method using the transitivity of equivalence class. The method tries to find two matched pairs that share the same entity so that it can infer another match between other two entities, because the matching is done by crowdsourcing and requires a huge monetary cost. Our work is different from it in the way to use the transitivity of equivalence because we use it to detect errors in AI decisions assuming AI is the main matcher. Other methods (Li, et al., 2020), (Trabelsi, et al., 2022) use domain knowledge to improve accuracy in entity matching. Their approach is completely different from ours and complementary to each other.

Problem definition

In this section, we define the problem. Table 1 shows definitions of symbols in this paper.

We define a tuple of entities as $x = (s_i, s_j)$, set of all possible tuple of entities as $\mathcal{X} = \{(s_i, s_j) \in \mathcal{S}^2 \mid i < j\}$, a set of labels $\mathcal{Y} = \{Match, Unmatch\}$, and a set of all true data as $\mathcal{U}^n = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, where $n = {}_{|\mathcal{S}|}C_2$. $\mathcal{L}^m = \{(x_i, y_i)\}_{i=1}^m \subseteq \mathcal{U}^n$is the labelled training set with m samples(We assume that humans give answers with the platform such as crowdsourcing) (we give m and $m \ll n$). Our main goal is to design a query strategy $\mathcal{Q}: \mathcal{U}^n \to \mathcal{L}^m$ to train an entity matching model (matcher) $f \in \mathcal{F}$, $f: \mathcal{X} \to \mathcal{Y}$. The optimization problem can be expressed as follows:

$$\underset{\mathcal{L}^m \subseteq \mathcal{U}^n}{\operatorname{argmax}} \frac{1}{n} \sum_{(x,y) \in \mathcal{U}^n} \delta(f(x) = y \vee (x, y) \in \mathcal{L}^m) \tag{1}$$

where δ is the indicator function. The intuition of Eq. (1) is that, for all pairs of entities in a dataset, it is beneficial for the matcher to answer a matching for a pair x correctly, and the dataset $\mathcal{L}^m$ contains the correct label for a pair x. We assume the model returns a probability of $Match$ or $Ummatch$, and we note the probability of $Match$ that the matching model gives to a tuple (s_i, s_j) as $P(Match | s_i, s_j)$.

Symbol	Definition
$\mathcal{S} = \{s_i\}$	All entities in the dataset.
s_i	An entity.
$y \in \mathcal{Y} = \{Match, Unmatch\}$	Labels.
$x \in \mathcal{X} = \{(s_i, s_j) \in \mathcal{S} \times \mathcal{S} \mid i < j\}$	Possible tuples.
$\mathcal{U}^n \subset \mathcal{X} \times \mathcal{Y}$	A set of all possible pairs and their labels.
$\mathcal{L}^m \subset \mathcal{U}^n$	Labeled entity tuples for retraining.
$\mathcal{Q}: \mathcal{U}^n \to \mathcal{L}^m$	A sampling strategy for labeling.
$f \in \mathcal{F}$	Model (e.g., Bayesian inference).

Table 1. Symbol definitions

Inconsistency-driven sampling

Basic idea

The inconsistency-driven sampling samples the pairs of the entities asked to the human. Intuitively, with the current labelled set $\mathcal{L}^{m'}$ we compute $P(Match \mid s_i, s_j)$ for all pairs in $\mathcal{U}^n - \mathcal{L}^{m'}$, find pairs that cause inconsistency, choose m pairs, and re-train the parameters of matching model.

To find the inconsistent pairs from the inference result by the current matching model, we consider pairs in a triple of the entities. We call the matching of the three pairs inconsistent if they constitute impossible patterns under the transitivity of equivalence relation (such as "positive," "positive," and "negative," as shown in Fig. 2).

Using Bayesian inference, we can calculate the probability that each pair is a match. Given a triple $\Delta_{ijk} = (s_i, s_j, s_k)$, the probability that these three pairs cause inconsistency can be calculated as in Eq. (2).

$$Inconsistency(\Delta_{ijk}) = \sum_{(a,b,c)\in\{(i,j,k),(j,k,i),(k,i,j)\}} P(Match|s_a, s_b)P(Match|s_b, s_c)P(Unmatch|s_c, s_a) \quad (2)$$

In this study, when the *Inconsistency* is higher, we assume the matcher is getting more confused, and we prioritize the triple to ask humans to correct it.

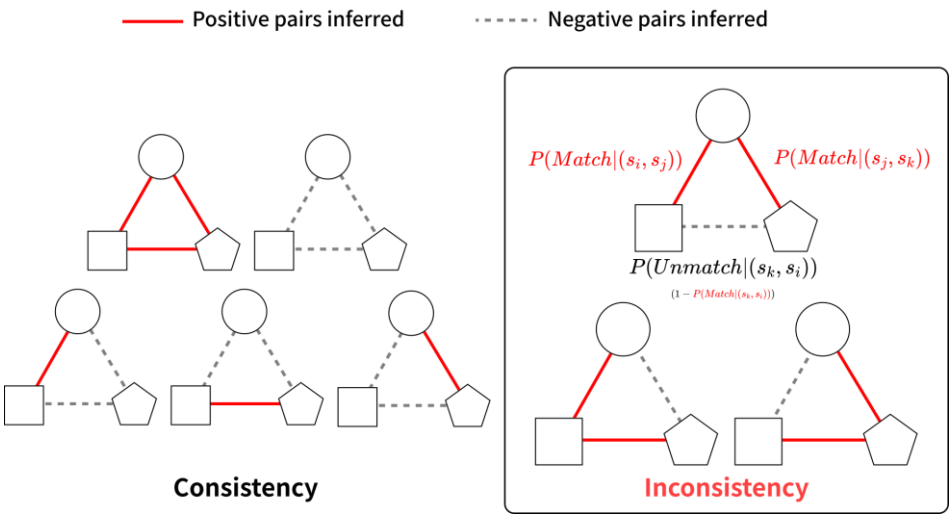


Figure 2. Inconsistency in focusing on the three data

Inconsistency-driven sampling algorithm

Algorithm 1 implements the idea of inconsistency-driven sampling. It takes a set of entity pairs with match probabilities $\hat{\mathcal{D}} = \{((s_i, s_j), p)\} \subset \mathcal{X} \times [0,1]$ and outputs m sample pairs taken from the given set. Note that the matching probability $p = P(Match|s_i, s_j)$. Note that the given set of entity pairs forms a graph, and that the triples that may cause the inconsistency in the graph are triangles. We define a set of triangles as

$$\mathcal{T} = \{\Delta_{ijk} \in \mathcal{S}^3 \mid ((s_i, s_j), p_1), ((s_j, s_k), p_2), ((s_k, s_i), p_3) \in \hat{\mathcal{D}}\} \quad (3)$$

The flow of inconsistency-driven sampling is as follows: First, we apply the blocking method for the set of entities, then sort the triangles in descending order of *Inconsistency*, and lastly, sample the $m/3$ triangles with the highest *Inconsistency* to sample m pairs of entities.

Algorithm 1 Inconsistency-driven sampling

Input: $\hat{\mathcal{D}}, m$
Output: A set of m pairs of entities
1: $\mathcal{T}' \leftarrow$ Sort $\Delta_{ijk} \in \mathcal{T}$ in descending order by $Inconsistency(\Delta_{ijk})$
2: **return** Triangles up to the $(m/3)$-th of $\mathcal{T}'$

Algorithm 1. Inconsistency-driven sampling

Experiments

We conducted an extensive set of experiments to address our research questions and evaluated the impact of sampling strategies on the f1 value of matchers on datasets from different domains.

Figure 3 overviews the overall experiment workflow. First, we apply a blocking technique to each of the three datasets for the experiment and generate training and evaluation datasets. Then, we execute human-in-the-loop entity matching iterations with each of the five sampling strategies. We explain the steps one by one.

Datasets and human settings

Datasets

We used three datasets taken from different domains (Table 2): **Persons** (Köpcke, et al., 2010), **Bibliorecords** (a private dataset supplied by public libraries in Japan), and **Music** (Köpcke, et al., 2010). Each dataset has a set of entities that has an attribute that stores cluster labels; if two have the same cluster label, they are matched entities, i.e., they represent the same entity. **Persons** and **Bibliorecords** are relatively clean datasets, while **Music** has many missing values and contains dirty attributes, such as having an album name in the (music) title attribute.

Domain (Language)	Attributes
Persons (Köpcke, et al., 2010) (English)	name, surname, suburb, postcode
Bibliorecords (Japanese)	title, author, publisher, date
Music (Köpcke, et al., 2010) (English)	artist, title, album, year, length

Table 2. Datasets from three different domains

Human settings

It is unavoidable for humans to make mistakes. In order to see the impact of errors on each sampling strategy in a systematic way, we took a simulation-based approach by implementing agents that serve as humans who give labels with a given accuracy; we examined three cases in terms of the accuracy of human labels: 100%, 95%, and 90%. The accuracy is, in other words, the percentage of noise labels in active learning, and we investigated the impact of noise labels on active learning (Wu, et al., 2022), (Younesian, et al., 2021).

Blocking and data preparation

For each of the three original datasets, we generated two data sets we use in the blocking phase and the human-in-the-loop entity matching iterations:

- the training dataset for the blocking phase and the Bayesian inference, and
- the evaluation dataset for evaluating sampling strategies to update the Bayesian inference.

The two datasets were constructed as follows and disjointed from each other.

Training dataset D^t: For each original dataset, we randomly choose a set of 15,000 positive pairs (the two cluster labels are the same) and 15,000 negative pairs (the labels are different from each other) from all pairs of entities in the original dataset. Then, every entity pair and their label form a triple contained in D^t.

Evaluation Dataset D^e: First, we randomly select clusters from each of the original datasets until each dataset contains about 2000 entities in total. Then, we generate a set of entity pairs with labels from the clusters.

Then, we constructed the dataset by applying a standard blocking technique based on metric learning to all pairs of the selected entities. The blocking technique chose only pairs of entities

that were closer than the threshold after metric learning on the distance among entities with the training dataset.

Table 3 shows the statistics of the evaluation datasets constructed this way. Note that the blocking does not work in favour of the inconsistency-driven sampling because the blocking is done based on the metric learning result, and it removed some of the potentially matched pairs that can affect the inconsistency-based sampling (i.e., the recall of the matches is less than 1).

Domain (Language)	#Entities	#Pairs	#Matches	Recall of matches of the blocking
Persons	2001	48,835	1379	0.848
Bibliorecords	2001	66,911	1594	0.840
Music	2002	196,414	1613	0.739

Table 3. Statistics of the evaluation datasets after blocking

Human-in-the-loop entity matching iterations

Algorithm 2 gives the concrete steps in the human-in-the-loop entity matching iterations in the experiment workflow. First, we conduct the Bayesian inference with $D^{\{t\}}$ for $D^{\{e\}}$. Then, in the iteration, it chooses samples for which it obtains human labels and uses the obtained result to update the inference. We chose Bayesian inference because it is a simple matcher that satisfies the requirement for the application of our inconsistency-driven sampling: it outputs a matching probability for a pair of entities. Note that our research question is not the performance of a particular matcher, and we can use this Bayesian inference matcher without loss of generality.

Algorithm 2 Human-in-the-loop entity matching iterations for the experiment

Input: Training Data $\mathcal{D}^t \subset \mathcal{U}^n$, Evaluation Data $\mathcal{D}^e \subset \mathcal{U}^n - \mathcal{D}^t$
Output: (None)
1: **for** $k = 1$ to 10 **do**
2: Conduct the Bayesian inference with $\mathcal{D}^t$ for $\mathcal{D}^e$ to output $\hat{\mathcal{D}}^e$
3: Calculate F1 value for $\hat{\mathcal{D}}^e$ (with $\mathcal{D}^e$)
4: Choose m pairs from $\hat{\mathcal{D}}^e$ using **Algorithm 1**
5: Obtain the human label for m pairs and add the data to $\mathcal{D}^t$
6: Update the matcher with m pairs
7: **end for**

Algorithm 2. Human-in-the-loop entity matching iterations for the experiment

Matcher implementation

The workflow implements a Bayesian inference-based matcher. The model infers the matching by considering the similarity measures between the given two entities. Our model requires the four similarities, which are the basic similarity measures for texts shown in Table 4.

l	Indicator
1	FastText vector (Bojanowski, et al., 2017)
2	Jaro-winkler (Winkler, et al., 1990)
3	Levenshtein (Levenshtein, et al., 1966)
4	Gestalt Pattern Matching (Virtanen, et al., 2020)

Table 4. Statistics of the evaluation datasets after blocking

The probabilities obtained from these probability density functions can be integrated as in Eq. (4) to estimate the matching probability of the pair.

$$P(\,Match \mid s_i, s_j\,) = \frac{\prod_{l=1}^{n} p(\,z_{i,j}^{(l)} \mid Match\,)P(Match)}{\sum_{Class\in\{Match,Unmatch\}} \prod_{l=1}^{n} p(\,z_{i,j}^{(l)} \mid Class\,)P(Class)} \tag{4}$$

Note that $z_{i,j} = \left(z_{i,j}^{(1)}, \cdots, z_{i,j}^{(n)}\right)^{\top}$ is a vector based on the similarity between entities s_i and s_j. The parameters for constructing the probability p are obtained by fitting a probability density function based on the mixed Gaussian distribution (Dempster, et al., 1977), (Pedregosa, et al., 2011) to the training data (Fig. 4).

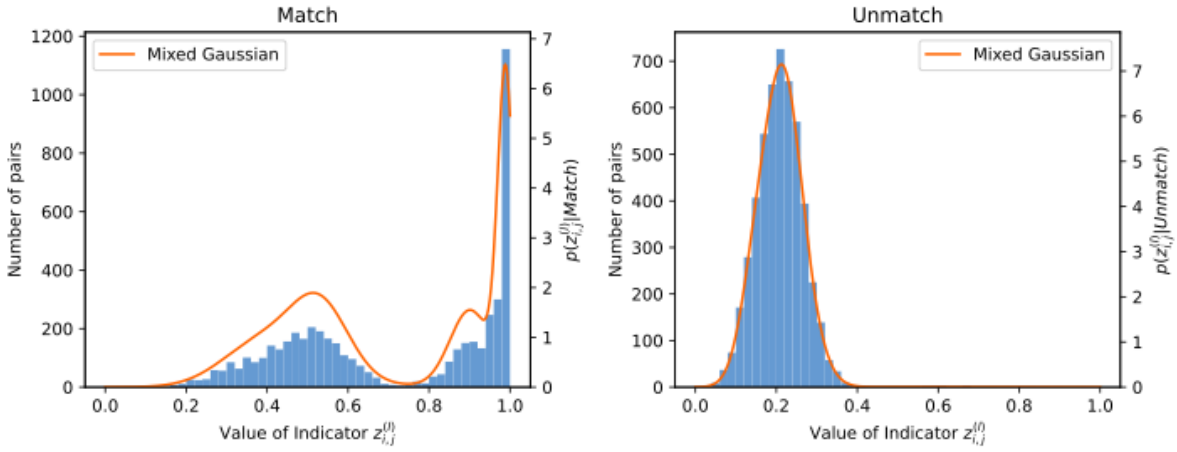


Figure 4. Fitting probability density function

Sampling strategies

We used the following five sampling strategies. Note that uncertainty sampling and query-by-committee sampling are *model-based* strategies while diversity sampling and random sampling are *model-free* strategies.

1. **Inconsistency-driven sampling**. This is the method we proposed in which pairs whose estimated labels lead to an inconsistency.

2. **Uncertainty sampling (model-based)** (Settles, 2010). In Bayesian inference, we compute the confidence value for each pair and chose those pairs that are not clearly positive or negative. Specifically, we use Eq. (5) for the confidence value and choose the pair with the least confidence value.

$$Priority = 1 - \left|P\left(Match|s_i, s_j\right) - 0.5\right| \tag{5}$$

3. **Query-by-committee sampling (model-based)** (Settles, 2010). Query-by-committee is a method of choosing pairs by aggregating the results of multiple indicators. In this experiment, we used two indicators that computed positive and negative scores for each indicator, which were calculated and sampled from antagonistic pairs. Specifically, sample in order of increasing value of Eq. (8).

$$Positive = \prod_{l=1}^{n} p\left(z_{i,j}^{(l)}|Match\right)P(Match) \tag{6}$$

$$Negative = \prod_{l=1}^{n} p\left(z_{i,j}^{(l)}|Unmatch\right)P(Unmatch) \tag{7}$$

$$Priority = 1 - (Positive + Negative) \tag{8}$$

4. **Diversity sampling (model-free)** (O'Neill, et al., 2017). This method selects distant pairs from those that have already been labelled by humans. The distance is calculated using the embedded representation of FastText (Bojanowski, et al., 2017). The first iteration is $\mathcal{L}^m = \emptyset$, so random sampling is executed only in the iteration.

5. **Random sampling (model-free)**. This method randomly selects the candidate pairs. We used the random function from Python's random module.

Other settings

Prior distribution for Bayesian inference. The prior distribution was set to $P(Match) = 0.1$ and $P(Unmatch) = 0.9$.

Batch sampling. We adopted a batch sampling scheme to reduce the number of inference updates; in each iteration, we choose m samples (instead of choosing one sample) and obtain their human labels before each inference update. We set $m = 300$. We need to be careful when dealing with inconsistency-driven sampling in the batch because we may obtain more than one human label for the same pair if it appears in different sets of inconsistent triangles. We solved the duplication by majority vote.

Languages and libraries. These algorithms were implemented by Python3, using the modules Tensorflow (Abadi et al., 2015) for metric learner construction, Cupy (Okuta et al., 2017) and Faiss (Johnson et al., 2019) for blocking and indexing, and Scipy (Virtanen, et al., 2020) and Scikit-Learn (Pedregosa, et al., 2011) for Bayesian inference probability density function manipulation.

Results

Difference of sampling distributions

Fig. 5 shows how different the five sampling strategies are from each other in terms of chosen samples. The horizontal axis represents the match probability determined by the inference, where the closer the match probability is to 1, the more likely the pair is to be matched, while the closer it is to 0, the more unlikely. The vertical axis represents the number of chosen pairs in the sampling strategy (in the log scale). The higher orange intensity means that they are chosen in earlier iterations.

The result clearly shows that the strategies are remarkably different in terms of chosen samples. Inconsistency-driven sampling tends to choose pairs that are highly likely or unlikely to be positive, while uncertainty sampling constantly chooses pairs in the middle. Query-by-committee, diversity, and random sampling choose pairs from a wide range.

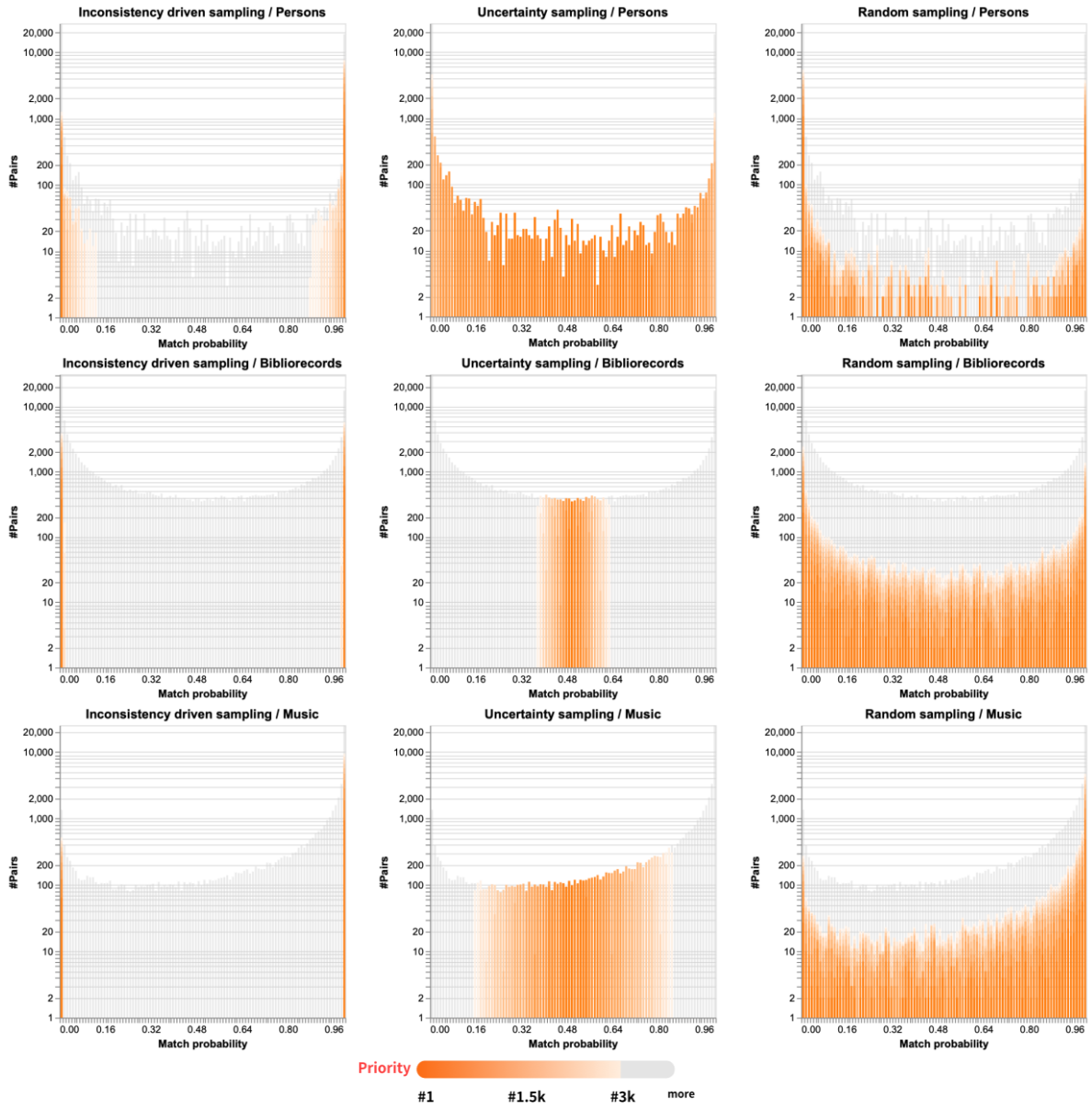


Figure 5. Analysis: Distribution of match probability for each sampling strategy

Effects on matcher performance

Fig. 6 shows the effects of each sampling strategy on the matcher performance in the iterations. The X-axis is the number of iterations. The Y-axis is the F1 value for the output of Bayesian Inference and human labels after each iteration for the evaluation dataset. Each line represents a sampling strategy.

The aim of seeing the figure is to identify the sweet spots of each sampling strategy. Note that the performance of matchers (to be used with sampling strategies) for **Persons**, **Bibliorecords**, and **Music** are very different (very high, moderate, and very low, respectively). The result suggests the following. First, inconsistency-driven sampling is effective when the f1 value of the matcher is high (i.e., higher than 0.8), while uncertainty sampling performs the best, especially for lower-quality matchers. Second, when the f1 value is high, the performance inconsistency-driven is stable, while other sampling strategies are directly influenced when the accuracy of human labels becomes lower. The results are reasonable for the following reasons: if the f1 value of matchers is not high,

it is difficult to identify inconsistency correctly. Third, the inconsistency-driven approach can identify inconsistencies even for newly obtained human labels, while other strategies take human inputs as oracles, even if they are incorrect.

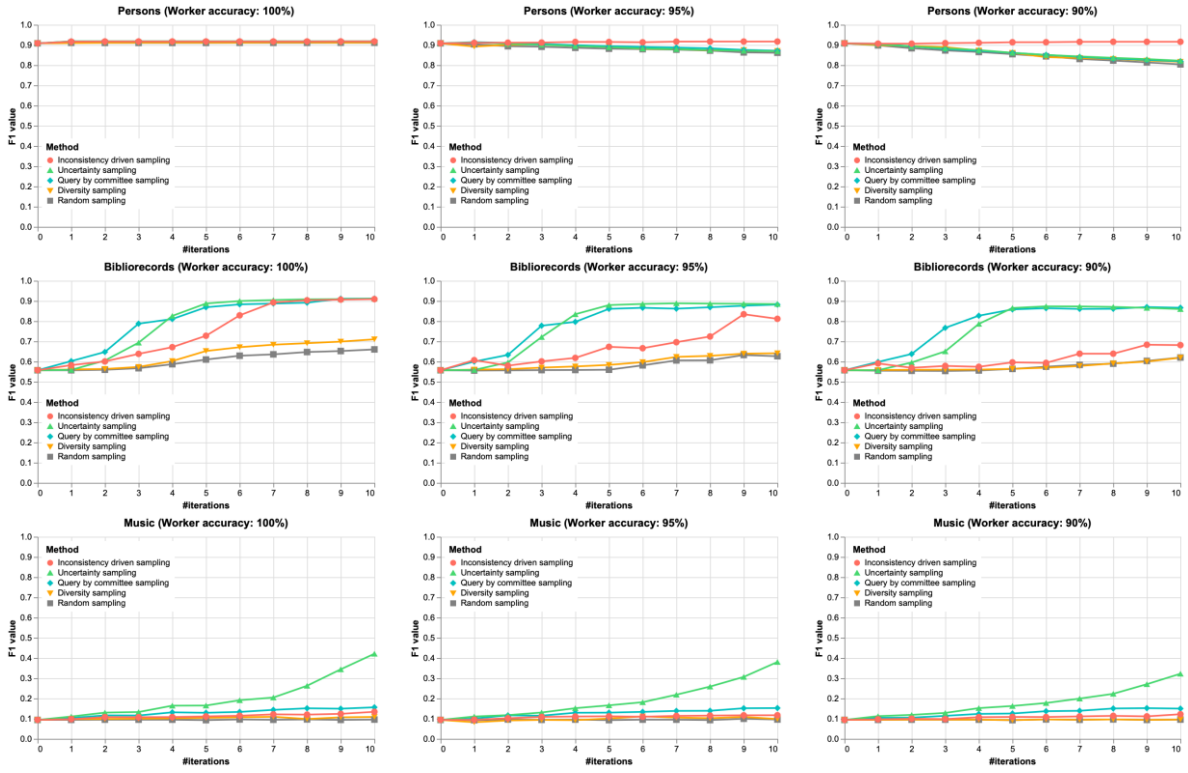


Figure 6. Experiment: Comparison of f1 value for different sampling strategies

Hybrid strategy

The results so far showed that (1) the inconsistency-based sampling works better when the inference f1 value is high while uncertainty sampling works better otherwise and that (2) the chosen pairs of the strategies are completely different to each other. They suggest that it is worth considering a hybrid strategy. Since it is difficult to set the threshold to switch strategies, we considered a simple hybrid strategy that switches uncertainty and inconsistency-driven strategies in each iteration.

Fig. 7 shows the results of applying the hybrid strategy to the three datasets with 90%-accurate human labels. The result shows that the hybrid strategy performs well when the inference f1 value is moderate or higher, which covers many practical situations. On the other hand, uncertainty sampling still works better with the inference with an extremely low f1 value.

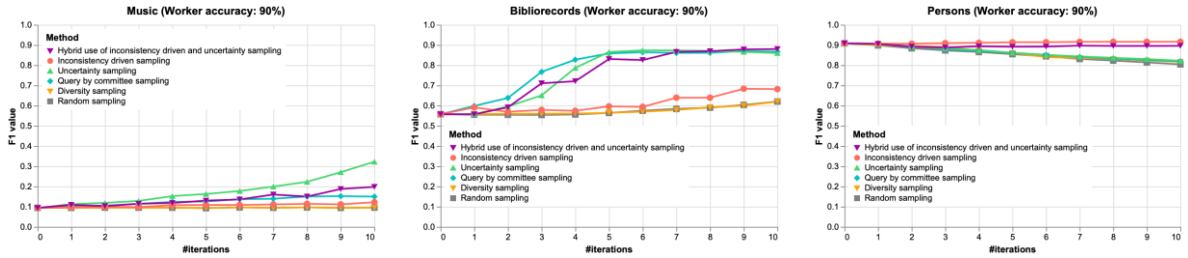


Figure 7. Experiment: comparison of f1 value for different sampling strategies

Conclusion

This paper addressed an inconsistency-based sampling strategy to deal with the *unknown-unknown problem* in active learning for entity matching. The method asks humans to resolve the entities when we find inconsistency in the transitivity property. This paper implemented a human-in-the-loop entity matching framework with this sampling strategy with similarity-based blocking method and Bayesian inference. It also explained the result of an extensive set of experiments that reveals how and when the method is effective. The result showed that the inconsistency-based sampling selects very different entity pairs compared to other sampling strategies and that a simple hybrid strategy performs well in many practical situations. Future work includes the interaction of the sampling strategy and matcher implementations. For example, the sampling result may suggest switching to other types of matchers.

Acknowledgements

This work was supported in part by JST CREST(JPMJCR22M2), Grants-in-Aid for Scientific Research (22H00508, 21H03552).

About the authors

Hiroyoshi Ito is an Assistant Professor at Institute of Library, Information and Media Science, University of Tsukuba. He can be contacted at: ito@slis.tsukuba.ac.jp

Takahiro Koizumi is a master's student at Graduate School of Comprehensive Human Sciences, University of Tsukuba. He can be contacted at: takahiro.koizumi.2022b@gmail.com

Ryuji Yoshimoto is an Engineer at CARLIL Inc. He can be contacted at: ryuuji@calil.jp

Yukihiro Fukushima is an Associate Professor at Faculty of Letters, Keio University. He can be contacted at: fukusima-y@keio.jp

Takashi Harada is a Professor at Center for License and Qualification, Doshisha University. He can be contacted at: ushi@slis.doshisha.ac.jp

Atsuyuki Morishima is a Professor at Institute of Library, Information and Media Science, University of Tsukuba. He can be contacted at: mori@slis.tsukuba.ac.jp

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), http://tensorflow.org/, software available from tensorflow.org

Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S.E., Widom, J. (2009). Swoosh: a generic approach to entity resolution. The VLDB Journal 18, 255–276.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., Stefanidis, K. (2020). An overview of end-to-end entity resolution for big data. ACM Comput. Surv. 53(6). https://doi.org/10.1145/3418896

Chung, Y., Haas, P.J., Upfal, E., Kraska, T. (2019). Unknown examples & machine learning model generalization.

Cohen, W.W., Ravikumar, P., Fienberg, S.E., et al. (2003). A comparison of string dis- tance metrics for name-matching tasks. In: IIWeb. vol. 3, pp. 73–78.

Das, S., G.C., P.S., Doan, A., Naughton, J.F., Krishnan, G., Deep, R., Arcaute, E., Raghavendra, V., Park, Y. (2017). Falcon: Scaling up hands-off crowdsourced entity matching to build cloud services. In: Proceedings of the 2017 ACM International Conference on Management of Data. pp. 1431–1446. https://doi.org/10.1145/ 3035918.3035960

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) 39(1), 1–22.

Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., Tang, N. (2018). Distributed representations of tuples for entity resolution. Proc. VLDB Endow. 11(11), 1454–1467. https://doi.org/10.14778/3236187.3236198

Gokhale, C., Das, S., Doan, A., Naughton, J.F., Rampalli, N., Shavlik, J., Zhu, X. (2014). Corleone: Hands-off crowdsourcing for entity matching. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. pp. 601–612. https://doi.org/10.1145/2588555.2588576

Govind, Y., Paulson, E., Nagarajan, P., C., P.S.G., Doan, A., Park, Y., Fung, G.M., Conathan, D., Carter, M., Sun, M. (2018). Cloudmatcher: A hands-off cloud/crowd service for entity matching. Proc. VLDB Endow. 11(12), 2042–2045. https://doi.org/10.14778/3229863.3236255

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. Journal of the American Statistical Association 84(406), 414. https://doi.org/10.2307/2289924

Johnson, J., Douze, M., Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data 7(3), 535–547.

Köpcke, H., Thor, A., Rahm, E. (2010). Evaluation of entity resolution ap- proaches on real-world match problems. Proc. VLDB Endow. 3(1-2), 484–493. https://doi.org/10.14778/1920841.1920904

Levenshtein, V.I., et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710.

Li, G. (2017). Human-in-the-loop data integration. VLDB Endowment 10(12), 2006–201.

Li, Y., Li, J., Suhara, Y., Doan, A., Tan, W.C. (2020). Deep entity matching with pre-trained language models. Proceedings of the VLDB Endowment 14(1), 50–60. https://doi.org/10.14778/3421424.3421431

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., Raghavendra, V. (2018). Deep learning for entity matching: A design space exploration. In: Proceedings of the 2018 International Conference on Management of Data. pp. 19–34. https://doi.org/10.1145/3183713.3196926

Nielsen, R.D. (2018). Introduction to machine learning for digital library applications. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. p. 421–422. https://doi.org/10.1145/3197026.3201780

Okuta, R., Unno, Y., Nishino, D., Hido, S., Loomis, C. (2017). Cupy:Anumpy-compatible library for nvidia gpu calculations. In: Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural In- formation Processing Systems (NIPS), http://learningsys.org/nips17/assets/papers/paper_16.pdf

O'Neill, J., Delany, S., MacNamee, B. (2017). Model-Free and Model-Based Active Learning for Regression, vol. 513, pp. 375–386. https://doi.org/10.1007/ 978-3-319-46562-3_24

Osawa, N., Ito, H., Fukushima, Y., Harada, T., Morishima, A. (2021). Bubble: A quality-aware human-in-the-loop entity matching framework. In: The 5th IEEE Workshop on Human-in-the-loop Methods and Future of Work in Big-Data (IEEE HMData2021). pp. 3557–3565. https://doi.org/ 10.1109/BigData52589.2021.9672002

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peeters, R., Bizer, C. (2022). Supervised contrastive learning for product matching. In: Companion Proceedings of the Web Conference 2022. pp. 248–251 https://doi.org/10.1145/3487553.3524254

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M. (1995). Okapi at trec-3, https://www.microsoft.com/en-us/research/ publication/okapi-at-trec-3/

Settles, B. (2010). Active learning literature survey. In: Active Learning Literature Survey. University of Wisconsin-Madison, https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf

Takashi, H., Yukihiro, F., Sho, S., Misato, T., Ryuji, Y., Atsuyuki, M. (2019). Advancement of bibliographic identification using a crowdsourcing system. Proceedings of the 9th Asia-Pacific Conference on Library & Information Education and Practice (A-LIEP 2019) pp. 71–82.

Trabelsi, M., Heflin, J., Cao, J. (2022). Dame: Domain adaptation for matching entities.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Courna- peau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., Vander- Plas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P. (2020). SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Winkler, W. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. Proceedings of the Section on Survey Research Methods pp. 354–359.

Wu, M., Li, C., Yao, Z. (2022). Deep active learning for computer vision tasks: Methodologies, applications, and challenges. Applied Sciences 12(16) https://doi.org/10.3390/app12168103

Xu, X., Zhang, F., Niu, Z. (2008). An ontology-based query system for digital libraries. IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application. vol. 1, pp. 222–226. https://doi.org/10.1109/PACIIA.2008. 360

Yao, D., Gu, Y., Cong, G., Jin, H., Lv, X. (2022). Entity resolution with hierarchical graph attention networks. In: Proceedings of the 2022 International Conference on Man- agement of Data. pp. 429–442. https://doi.org/10.1145/3514221.3517872

Younesian, T., Zhao, Z., Ghiassi, A., Birke, R., Chen, L.Y. (2021). Qactor: Active learning on noisy labels. In: Balasubramanian, V.N., Tsang, I. (eds.) Proceedings of The 13th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 157, pp. 548–563., https://proceedings. mlr.press/v157/younesian21a.html

Zhu, Y., Liu, H., Wu, Z., Du, Y. (2020). Relation-aware neighborhood matching model for entity alignment. https://arxiv.org/abs/2012.08128

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

The effects of anthropomorphic framing on senior news consumers' attitudes towards health AI systems: a mediation of psychological distance

Yanrun Xu and Tingting Jiang

DOI: https://doi.org/10.47989/ir30iConf47128

Abstract

Introduction. Mass media plays a critical role in demonstrating the advancements of AI technologies to the public. Anthropomorphic framing, a verbal technique involving attributing human characteristics to non-human entities, has been increasingly adopted to describe various health AI systems, but there still lacks empirical evidence regarding its effectiveness in enhancing AI acceptance among senior news consumers.

Method. This study conducted a controlled experiment based on a single-factor (human- vs. machine-like framing) within-subject design. 37 senior participants were asked to watch short-form video news discussing health AI systems, and their attitudes towards, intentions to use, and psychological distance from health AI systems were measured using appropriate scales.

Results. Describing a health AI system like a human being elicited more positive attitudes towards AI among the elderly, which further increased their intentions to use AI. The positive effect of anthropomorphic framing on AI attitudes was mediated by psychological distance.

Conclusions. Anthropomorphic framing has great potential in the AI literacy education for older adults. This study not only extends research on AI anthropomorphism but also provides practical implications for media outlets to create persuasive news content.

Introduction

The world's population is ageing (United Nations, 2022). This implies a greater need for social and technological resources to address healthcare-related issues for the elderly (Huang & Huang, 2020). Artificial intelligence (AI) offers new opportunities for older adult healthcare practices, aiding in the enhancement of elderly patient care, streamlining medical processes, and improving overall health outcomes (O'Connor, 2022). However, there are significant barriers to promoting medical AI among seniors. A survey of adults over the age of 50 found that 49% of participants felt uneasy about using AI systems for diagnosing medical issues (NORC, 2023). Despite acknowledging the increasing prevalence of AI, elderly patients show much higher skepticism of the technology and remain reluctant to allow AI to handle their primary care and therapy (McKnights, 2023). Thus, enhancing older adults' attitudes toward and willingness to use medical and healthcare AI systems has become a pressing issue.

News is a crucial source for older adults to learn about AI and its applications in healthcare. However, news media are not neutral observers and they shape perceptions, understanding, and attitudes towards technology through news framing (Groves et al., 2016; Nguyen & Hekman, 2024). These frames have been noted as an important mental shortcut that influences how people think about emerging technologies of which they lack direct experiences and clear understanding (Choi, 2024). In recent years, news reports on the application of AI in the healthcare sector have tended to adopt an '***anthropomorphic framing***' to describe AI, essentially personifying AI as a human being. For instance, a *Wall Street Journal* article introducing a medical AI system was titled '*The AI Doctor Will See You Now*' (Mims, 2018). A qualitative analysis of 365 coverages related to AI in medicine, published between 1980 and 2019, found that anthropomorphism has gradually become one of the most commonly used frames in AI-related news reporting (Bunz & Braghieri, 2022).

Prior studies from human-computer interaction and psychology provided conflicting evidence regarding the impact of anthropomorphic framing of AI on user perceptions in non-healthcare contexts (Barone et al., 2024; Kopp et al., 2022; Roesler, 2023). Moreover, these related studies have primarily focused on younger adults, while the effects of anthropomorphic news framing of AI on older adults may significantly differ from those on younger people. Thus, how to describe healthcare AI in news reports is a crucial consideration for practitioners when writing news articles.

This study specifically focused on older adults, a key demographic for AI healthcare services, and explored whether and how anthropomorphic framing in healthcare AI news influences the older adults' attitudes toward AI. Drawing on the related literature of anthropomorphism, this study posits that the anthropomorphic news framing of AI can influence the older adults' attitudes through the pathway of psychological distance. When AI is endowed with human-like attributes, people perceive a higher similarity with AI, thereby reducing the psychological distance between themselves and AI (Ahn et al., 2021).

Currently, news media emphasize the new opportunities that AI brings to the healthcare services for older adults. Therefore, this study examined the impact of anthropomorphic news framing on older adults' attitudes toward AI and its underlying mechanisms in the context of positive news. The conceptual framework can be seen in **Figure 1**. This study extends research on AI anthropomorphism by shifting the focus from '*designed social cues*' to '*described social cues*'. The former refers to social cues embedded in AI systems, allowing participants to directly interact with the system using these cues. In contrast, the latter are presented in textual form, where participants cannot interact with the AI directly but instead imagine it based on the textual descriptions provided. Additionally, this study offers valuable implications for journalists on how to write news about medical and healthcare AI systems. The specific research questions are as follows:

- **RQ1.** Does AI anthropomorphic news framing influence the older adults' attitudes toward health AI system?
- **RQ2.** Does psychological distance mediate the impact of the AI anthropomorphism framing on the older adults' attitudes toward health AI system?
- **RQ3.** How do older adults' attitudes toward health AI system influence their intention to use it?

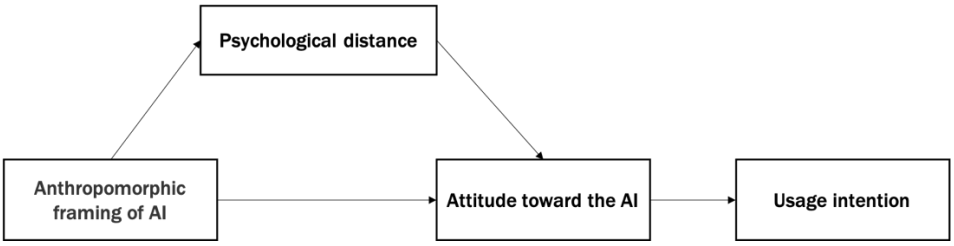


Figure 1 The conceptual framework

Literature review

Anthropomorphic framing of AI

The research focus on AI anthropomorphism has gradually shifted from the anthropomorphic design of AI social cues to anthropomorphic descriptions, which is the use of anthropomorphic rhetoric to describe AI features in text. Players read written instructions about a robot and the game rules before collaborating with the robot to complete a math game. The purpose of the human-like framed robot was perceived more positively after experiencing a comprehensible failure than the machine-like framed robot (Roesler, 2023). However, another study that conducted the same experiment found opposite results, indicating that participants were more willing to donate to a robot described using machine-like framing (Onnasch & Roesler, 2019). Moreover, a collaborative robot described with human-like framing significantly increased company employees' trust in their robot colleagues, but this effect only existed when the human-robot relationship was perceived as collaborative. When employees perceived the relationship with their robot colleagues as competitive, the human-like framing actually reduced their trust in the robot colleagues (Kopp et al., 2022). These mixed findings suggest a need to explore the consequences of anthropomorphic news framing of healthcare AI.

Older adults' attitudes towards AI

Older adults are generally considered to be technologically inadequate. They tend to adhere more strongly to old habits and resist changes brought about by new technologies (Broady et al., 2010). Evidence for an effect of AI anthropomorphic design on the elderly's attitudes toward AI is also scarce and somewhat inconsistent. For instance, when robots provide care services to the elderly, older adults tend to exhibit more positive attitudes towards robots with a lower degree of anthropomorphic appearance (Lehmann et al., 2021). Conversely, in the context of robots depicted in movies, the higher the degree of anthropomorphic appearance, the more positive the attitudes of older adult viewers towards the robots (Sundar et al., 2016). Therefore, it is hard to determine how AI anthropomorphic framing influences the attitudes of older adults towards AI.

Mediator: psychological distance

Psychological distance refers to the subjective experience of something or someone being close to or far from the self, here, and now (Trope & Liberman, 2010). Previous research has also provided empirical evidence for the mediating role of psychological distance between AI anthropomorphism and user responses. Anthropomorphic AI shopping assistants can strengthen the relationship

between consumers and brands by reducing the psychological distance between consumers and AI (Jham et al., 2023). Users exhibit more positive attitudes towards highly anthropomorphic AI assistants, and this effect is mediated by psychological distance (Li & Sung, 2021). In chatbot health counseling, users experience a closer psychological distance with highly anthropomorphic chatbots compared to low-anthropomorphic ones, making them more willing to follow the mental health advice provided by the highly anthropomorphic chatbots (Park et al., 2023). Therefore, it can be inferred that, compared to machine-like framing, human-like framing can reduce the psychological distance between older adults and AI, thereby leading to more positive attitudes towards AI among older adults.

Methods

This study conducted a controlled experiment to manipulated anthropomorphic framing in a single-factor (human-like framing vs. machine-like framing) within-subjects design. Senior participants were asked to watch short-form videos discussing health AI systems. These videos introduced the main functions and application advantages of AI-aided surgical robots and guided robots. These two health AI systems have been widely promoted and implemented in hospitals globally, yet older adults generally remain unfamiliar with them.

Participants

Participants for this experiment were recruited from several residential communities in China. Eligibility criteria included: (1) 55 years of age or older; (2) normal or corrected-to-normal visual acuity; (3) no hearing impairments; (4) good cognitive status; and (5) capable of unobstructed communication. A total of 37 individuals met all the requirements and expressed interest in participating. Since 5 individuals were excluded for failing to complete certain steps of the experiment, the final sample comprised of 32 participants, including 17 males and 15 females aged between 55 and 89 (M_{age} = 59.38). The participants were distributed as follows in terms of their educational background: 15.63% (N = 5) had a lower than a high school education, 37.5% (N = 12) had completed high school, and 46.88% (N = 5) held a bachelor's degree or higher.

Materials

This study collaborated with a professional video production team to create short videos featuring health AI. Information was conveyed through voiceovers and keywords displayed in the video. The difference between the human-like framing and machine-like framing videos was only in the script, while keeping all other aspects of the videos identical. This experiment used two real news reports about surgical robots (People.cn, 2023) and guided robots (People.cn, 2018) as prototypes, with appropriate content reductions and adaptations. Specific information such as surgery names, disease names, and hospital names were omitted to avoid interference with the experimental results.

Manipulation of anthropomorphic framing. The anthropomorphic description of AI primarily takes two forms (Bunz & Braghieri, 2022): (1) Name anthropomorphism, which involves giving the AI a human-like name; (2) Behavior anthropomorphism, which involves describing the AI's actions as if it were a human. Therefore, this study manipulated the anthropomorphic framing by altering the robot's name (e.g., mechanical device vs. robot doctor), body parts (e.g., storage system vs. brain), and the description of behaviors (e.g., the control console of the surgical robot can present many minute details vs. the surgeon's eyes can see many minute details). A total of four health AI short videos were produced, featuring the following combinations: human-like framing with a surgical robot, machine-like framing with a surgical robot, human-like framing with a guided robot, and machine-like framing with a guided robot. All videos were kept to a duration of approximately one minute.

An additional 18 participants (10 males, 8 females, M_{age} = 61.28) who met the experimental requirements were recruited to test whether the manipulation of the anthropomorphic framing was successful. Each participant watched the four videos in a random order. After watching each video, they were asked to rate the degree of AI anthropomorphism on a 7-point semantic differential scale (Li et al., 2023). The results of the paired samples t-test show that participants perceived the human-like framed surgical robot as significantly more anthropomorphic than the machine-like framed surgical robot ($M_{human-like}$ = 12.06, $M_{machine-like}$ = 9.17, p = 0.039 < 0.05). Similarly, participants perceived the human-like framed guided robot as significantly more anthropomorphic than the machine-like framed guided robot ($M_{human-like}$ = 12.61, $M_{machine-like}$ = 9.11, p = 0.018 < 0.05). There is no significant difference in the perceived anthropomorphism between the human-like framed surgical robot and the human-like framed guided robot (p = 0.593 > 0.05). Similarly, there is no significant difference in the perceived anthropomorphism between the machine-like framed surgical robot and the machine-like framed guided robot (p = 0.963 > 0.05). These results indicate that the manipulation of the anthropomorphic framing was successful.

Tasks and procedures

This experiment was conducted offline using a tablet to play the videos. Given that many older adults had difficulty accurately comprehend the questionnaire items, researcher verbally presented the questions in a more comprehensible manner and recorded their responses accordingly. To eliminate any influence of the questioning way on participants' answers, data collection was performed by a single researcher.

The specific procedures are as follows: first, demographics of the participants, including gender and age, was collected. Then, each participant was required to randomly watch one human-like framing video and one machine-like framing video, with different types of AI in each video (e.g., human-like framing and surgical robot, machine-like framing, and guided robot). After watching each video, participants filled out a questionnaire measuring their perception of AI anthropomorphism (Li et al., 2023), psychological distance (Aron et al., 1992), attitudes towards the AI (Li & Sung, 2021), and usage intention (Liu et al., 2022).

All participants signed a consent form and a confidentiality agreement prior to the experiment and were compensated with 20 CNY (approximately 3 USD) upon completion. This experiment was granted approval from the University Research Ethics Committee and adhered strictly to the general ethical guidelines.

Results

Reliability analysis

Cronbach's alpha was used to access the reliability of the scales used in this study. The Cronbach's alpha value of anthropomorphism (Cronbach's α = 0.868), attitude toward the AI (Cronbach's α = 0.864), and usage intention (Cronbach's α = 0.942) all exceed the recommended threshold value of 0.7, which suggests high construct reliability. Psychological distance is measured using a single item, thus there is no need for a reliability analysis.

Manipulation check

The results of the paired samples t-test showed that participants perceived the human-like framed health AI as significantly more anthropomorphic than the machine-like framed health AI ($M_{human-like}$ = 21.94, $M_{machine-like}$ = 20.16, p < 0.05). Therefore, the anthropomorphic framing was successfully manipulated in this experiment.

The effects of anthropomorphic framing on AI attitude

The repeated-measures ANOVA revealed a significant effect of anthropomorphic framing on senior participants' attitudes towards health AI (F (1, 31) = 5.244, p = 0.029 < 0.05). Specifically,

compared to machine-like framed health AI (M = 17.16, SD = 3.79), participants had more positive attitudes towards human-like framed health AI (M = 17.97, SD = 2.99).

Mediation analysis of psychological distance

To explore whether the effect of anthropomorphic framing on AI attitude is mediated by psychological distance, this study used PROCESS Model 4 (parallel mediation model) with a bootstrapped sample of 5000 (Hayes, 2015). As in **Figure 2**, both the effect of anthropomorphic framing on psychological distance (b = -0.531, p < 0.05) and that of psychological distance on AI attitude (b = -0.427, p < 0.05) were significant. Thus, the effect of anthropomorphic framing on AI attitude was mediated by psychological distance (b = 0.774, Boot SE = 0.415, 95% Boot CI [0.083, 1.705]). Furthermore, the direct effect was not significant (b = 0.774, p = 0.962), suggesting complete mediation in which psychological distance fully explained the relationship between anthropomorphic news framing and AI attitude.

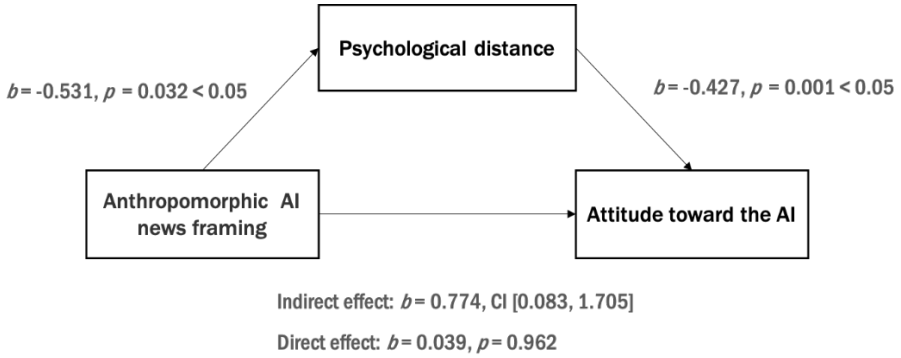


Figure 2. Mediation analysis results

The effects of AI attitude on usage intention

This study treated AI attitude as the independent variable in the regression analysis, with usage intention as the dependent variable, while age, gender, and education level were controlled for as covariates. The results showed that the attitudes of senior participants towards health AI significantly and positively influenced their willingness to use it (β = 0.656, p < 0.05). In other words, a positive attitude towards health AI can transform into a willingness to use it among the older adults.

Discussion

This study explored how anthropomorphic framing influences older adults' attitudes towards healthcare AI and the underlying mechanism. The major findings are summarized as follows.

For RQ1, this study found that older adults have a more favorable attitude towards healthcare AI in human-like framing messages. This finding extends previous human-AI interaction research based on the Computers are Social Actors Paradigm (CASA), demonstrating that whether humans interact directly with AI or indirectly (e.g., reading textual descriptions about AI, watching videos about AI), they exhibit more social perception and responses towards AI with more social cues.

For RQ2, it was found that the influence of anthropomorphic framing on older adults' AI attitudes is mediated by psychological distance. Compared to machine-like framing, human-like framing can signal low distance that reduce the psychological distance between older adults and AI, thereby improving their attitudes towards AI. This result supports previous research on psychological distance, demonstrating that psychological distance is an important mechanism between AI anthropomorphism and users' social responses (Jham et al., 2023; Li & Sung, 2021). However, unlike previous studies, this research focused on describes social cues of AI rather than designed social

cues. This subtle difference indicates that it is not the anthropomorphic form itself that changes users' psychological distance, but the social perception that anthropomorphism brings.

For RQ3, older adults' attitudes towards AI in the messages can positively influence their intention to use it. This finding basically echoes those of previous studies that focused on people's usage intention of healthcare or medical technologies (Bondzie - Micah et al., 2022; Lee & Lee, 2020). Although AI possesses higher autonomy and intelligence compared to ordinary medical devices, it essentially remains a tool for providing healthcare services. Human-AI interaction will continue to be primarily user-driven for the foreseeable future. Improving users' attitudes towards AI is critical to increasing their actual usage behavior.

Conclusions

This study aimed to explore the application of anthropomorphic framing in healthcare AI news. Results from a controlled experiment indicate that older adults exhibit more positive attitudes towards AI in human-like framing messages rather than machine-like framing messages. Such effect was mediated by psychological distance. Furthermore, older adults' positive attitudes towards health AI can transform into their willingness to use it.

The experimental research presented in this paper is a pioneering attempt to examine the roles of anthropomorphic framing in enhancing older adults' attitudes towards healthcare AI. Unlike most studies on anthropomorphic design of AI social cues, this study considers a relatively new form of anthropomorphism—anthropomorphic framing. The findings indicated that practitioners can adopt anthropomorphic descriptions of AI when the healthcare AI news targeting older adults.

Future research could explore the impact of different AI framings on audience attitudes and behavioral intentions. Moreover, considering personal characteristics, such as cultural background and gender, would provide a more comprehensive understanding of how described AI social cues influence individual perceptions and reactions.

Acknowledgements

This research has been made possible through the financial support of the National Natural Science Foundation of China under Grants no. 72074173.

About the authors

Tingting Jiang is a professor at the School of Information Management, Wuhan University, China. She received her Ph.D. from the University of Pittsburgh. Her current research interests include human-AI interaction, information behavior, user experience, and data analytics. She has published extensively in international journals, including *Government Information Quarterly*, *Information Processing and Management*, *Journal of Documentation*, *Internet Research*, and *International Journal of Medical Informatics*. She can be contacted at tij@whu.edu.cn.

Yanrun Xu is a Ph.D. candidate at the School of Information Management, Wuhan University, China. Her current research focuses on human-AI interaction and health information behavior. Her research has been published in *information processing and management* and iConference Proceedings, etc. She can be contacted at yarunx@whu.edu.cn.

References

Ahn, J., Kim, J., & Sung, Y. (2021). AI-powered recommendations: the roles of perceived similarity and psychological distance on persuasion. *International Journal of Advertising*, 40(8), 1366-1384. http://doi.org/10.1080/02650487.2021.1982529

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4), 596-612. http://doi.org/10.1037/0022-3514.63.4.596

Barone, A. M., Stagno, E., & Donato, C. (2024). Call it robot: anthropomorphic framing and failure of self-service technologies. *Journal of Services Marketing*, 38(3), 272-287. http://doi.org/10.1108/JSM-05-2023-0169

Bondzie - Micah, V., Qigui, S., Arkorful, V. E., Lugu, B. K., Bentum - Micah, G., & Ayi - Bonte, A. N. A. (2022). Predicting consumer intention to use electronic health service: An empirical structural equation modeling approach. *Journal of Public Affairs*, 22(4), e2677. https://doi.org/10.1002/pa.2677

Broady, T., Chan, A., & Caputi, P. (2010). Comparison of older and younger adults' attitudes towards and abilities with computers: Implications for training and learning. *British Journal of Educational Technology*, 41(3), 473-485. https://doi.org/10.1111/j.1467-8535.2008.00914.x

Bunz, M., & Braghieri, M. (2022). The AI doctor will see you now: assessing the framing of AI in news coverage. *AI & SOCIETY*, 37(1), 9-22. http://doi.org/10.1007/s00146-021-01145-9

Choi, S. (2024). Temporal Framing in Balanced News Coverage of Artificial Intelligence and Public Attitudes. *Mass Communication and Society*, 27(2), 384-405. http://doi.org/10.1080/15205436.2023.2248974

Groves, T., Figuerola, C. G., & Quintanilla, M. (2016). Ten years of science news: A longitudinal analysis of scientific culture in the Spanish digital press. *Public Underst Sci*, 25(6), 691-705. http://doi.org/10.1177/0963662515576864

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, 50(1), 1-22. http://doi.org/10.1080/00273171.2014.962683

Huang, T., & Huang, C. (2020). Elderly's acceptance of companion robots from the perspective of user factors. *Universal Access in the Information Society*, 19(4), 935-948. http://doi.org/10.1007/s10209-019-00692-9

Jham, V., Malhotra, G., & Sehgal, N. (2023). Consumer-brand relationships with AI anthropomorphic assistant: role of product usage barrier, psychological distance and trust. *The International Review of Retail, Distribution and Consumer Research*, 33(2), 117-133. https://doi.org/10.1080/09593969.2023.2178023

Kopp, T., Baumgartner, M., & Kinkel, S. (2022). How Linguistic Framing Affects Factory Workers' Initial Trust in Collaborative Robots: The Interplay Between Anthropomorphism and Technological Replacement. *International Journal of Human-Computer Studies*, 158, 102730. https://doi.org/10.1016/j.ijhcs.2021.102730

Lee, S. M., & Lee, D. (2020). Healthcare wearable devices: an analysis of key factors for continuous use intention. *Service Business*, 14(4), 503-531. https://doi.org/10.1007/s11628-020-00428-3

Lehmann, S., Ruf, E., & Misoch, S. (2021). *Emotions and Attitudes of Older Adults Toward Robots of Different Appearances and in Different Situations.* In Proceedings of the Information and Communication Technologies for Ageing Well and e-Health, 21-43.

Li, Q., Luximon, Y., & Zhang, J. (2023). The Influence of Anthropomorphic Cues on Patients' Perceived Anthropomorphism, Social Presence, Trust Building, and Acceptance of Health Care Conversational Agents: Within-Subject Web-Based Experiment. *J Med Internet Res*, 25, e44479. http://doi.org/10.2196/44479

Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User–AI assistant interactions. *Computers in Human Behavior*, 118, 106680. https://doi.org/10.1016/j.chb.2021.106680

Liu, Y.-l., Yan, W., Hu, B., Li, Z., & Lai, Y. L. (2022). Effects of personalization and source expertise on users' health beliefs and usage intention toward health chatbots: Evidence from an online experiment. *DIGITAL HEALTH*, 8, 20552076221129718. http://doi.org/10.1177/20552076221129718

McKnights (Producer). (2023). Seniors still skeptical of AI-led healthcare, new study shows. Retrieved from https://www.mcknightsseniorliving.com/news/seniors-still-skeptical-of-ai-led-healthcare-new-study-shows/

Mims, C. (2018). The AI Doctor Will See You Now. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/the-ai-doctor-will-see-you-now-1526817600

Nguyen, D., & Hekman, E. (2024). The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation. *AI & SOCIETY*, 39(2), 437-451. http://doi.org/10.1007/s00146-022-01511-1

NORC (Producer). (2023). Older Adults Express Mixed Views on Artificial Intelligence. Retrieved from https://www.norc.org/research/library/older-adults-express-mixed-views-artificial-intelligence.html

O'Connor, S. (2022). Artificial Intelligence for Older Adult Health: Opportunities for Advancing Gerontological Nursing Practice. *J Gerontol Nurs*, 48(12), 3-5. http://doi.org/10.3928/00989134-20221107-01

Onnasch, L., & Roesler, E. (2019). Anthropomorphizing Robots: The Effect of Framing in Human-Robot Collaboration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 1311-1315. http://doi.org/10.1177/1071181319631209

Park, G., Chung, J., & Lee, S. (2023). Human vs. machine-like representation in chatbot mental health counseling: the serial mediation of psychological distance and trust on compliance intention. *Current Psychology*, 1-12. https://doi.org/10.1007/s12144-023-04653-7

People. cn. (2018). The first intelligent medical guide robot is on duty at Mianyang Central Hospital in Sichuan province. Retrieved from http://health.people.com.cn/n1/2018/1011/c421589-30334783.html

People. cn. (Producer). (2023). Domestic robots move from laboratory to operating table. Retrieved from http://kpzg.people.com.cn/n1/2023/0914/c404214-40077496.html

Roesler, E. (2023). Anthropomorphic framing and failure comprehensibility influence different facets of trust towards industrial robots. *Front Robot AI*, 10, 1235017. http://doi.org/10.3389/frobt.2023.1235017

Sundar, S. S., Waddell, T. F., & Jung, E. H. (2016). *The Hollywood Robot Syndrome media effects on older adults' attitudes toward robots and adoption intentions*. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 343-350. http://doi.org/10.1109/HRI.2016.7451771

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological review*, 117(2), 440-463. http://doi.org/10.1037/a0018963

United Nations. (Producer). (2022). Ageing. Retrieved from https://www.un.org/en/global-issues/ageing

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Understanding entangled human-technology-world relations: use of intelligent voice assistants by older adults

Alisha Pradhan, Shaan Chopra, Pooja Upadhyay, Robin Brewer, and Amanda Lazar

DOI: https://doi.org/10.47989/ir30iConf47119

Abstract

Introduction. Emerging technologies like intelligent voice assistants or social robots can shape human relations with the world. To illustrate how an emerging technology mediates relations and shapes social practices in the context of aging, we present findings on use of voice assistants by older adults.

Method. We analysed interviews with 24 older adults by adopting a post-phenomenological perspective to examine how an emerging technology actively mediates relations between older individuals and their larger social world.

Results. Our findings surface the different types of relations that voice assistants mediate between older adults and their larger social world, unpacking how these relations shape social practices around what it means to give company to pets, to live alone, or to give and receive care.

Discussion. We discuss implications for understanding the mutually constitutive relations between older adults and the emerging technologies they use and opportunities in designing to support neglected relations, and accounting for nonhuman actors in technology and aging research.

Conclusion. We provide a preliminary understanding on how an emerging technology shapes social practices in later life. This understanding is crucial for aging and technology research, as several emerging technologies (e.g., social robots) target older adults, yet little is known about the relationships and discursive practices that shape their use.

Introduction

With emerging technologies such as, social robotics, intelligent assistants, virtual and augmented reality, technology is no longer a passive tool to be used by humans. Unlike the early days of computing, the influence of such technologies is not just limited to a single static computer screen, but actively permeates and shapes several aspects of our existence and relations. These technologies are ubiquitous, distributed, or merge the physical and the digital worlds, and thereby get woven into several aspects of our lives. Treating such technologies as passive tool that humans interact with, can fail to capture the complex ways in which they actively mediate human relations (Frauenberger, 2019). Understanding how these technologies mediate human relations with the world and how they shape social practices will not only yield new implications for design, but also help us attune to the power that these technologies have in shaping our society. To illustrate how an emerging technology actively mediates human relations and can shape social practices, in this work we present findings on use of AI-based intelligent voice assistants by older adults.

Home-based intelligent voice assistant (e.g., Alexa on Amazon Echo), also referred to as intelligent personal assistant, intelligent voice assistant or voice assistant, is a popular AI-based emerging technology that has received significant attention in context of aging-in-place— which involves the ability to live in one's own home and community safely, independently, and comfortably, regardless of age, income, or ability level (*CDC - Healthy Places - Healthy Places Terminology*, 2017). These voice assistants are being used by a growing number of older adults in their own homes (Carl Weinschenk, 2021; Federica Laricchia, 2022), installed in nursing homes and senior living communities (Emily Olsen, 2021; Shantanu Kurhekar, 2021), garnering interest as a commodity technology that can support aging-in-place. Research has noted how voice assistants can support independence and autonomy through features such as voice-controlled information access, controlling the home environment (e.g., through connected thermostats, lights, and blinds) (Pradhan et al., 2018). Voice assistants can also support other dimensions of aging in one's home such as access to leisure activities (playing music, games, jokes, stories), communication (hands-free calling), and managing routines (calendars, timers, alarms). As these technologies are being integrated into older adults' homes, researchers have investigated older adults' perceptions and use of voice assistants (Kim & Choudhury, 2021; Oewel et al., 2023; Pradhan et al., 2020), with a particular focus on how these technologies can support their health and well-being (Brewer et al., 2021; Brewer, 2022). These studies centre around older adults' attitudes, motivations, and practices of using these voice assistants at home (e.g., older adults' use this technology for accessing online information (Pradhan et al., 2020)).

However, centring research on solely human use and perceptions (in this case, of older adults) to understand technology use at home can miss some of the complex ways in which the technology, along with other nonhuman entities, shape human experiences. For example, in a family setting, individuals may discuss questions with the technology as a mediator, where the voice assistant plays an active role in altering the dynamics of conversation and relations at home (noted in works with younger adults and families (Beneteau et al., 2020; Porcheron et al., 2018)). While research has looked into how voice assistants can support social interactions for older adults, much of this work focuses on companionship through technology (Pradhan et al., 2019), and we know little about the other ways in which voice assistants can shape relations and interactions in older adults' homes. As individuals begin to adopt and live with such technologies in later life, there is an opportunity to understand the mutually constitutive relationship between individuals and technology. In this paper, we present findings from interviews with 24 older adults to answer the following research questions: How do home-based voice assistants mediate relations between older adults and their larger world at home (RQ1)? What are the different relations that constitute older adults, their larger world, and home-based voice assistants (RQ2)?

In this work, we unpack these salient relations constituting technology, older participants in our study, and their larger world. We analysed interviews with 24 older adults by adopting a post-phenomenological perspective. Our work contributes by providing a preliminary understanding and categorization of the salient relations mediated by intelligent voice assistants in older adults' homes. We further unpack how these mediations shape social practices around what it means to give company to pets, to live alone, or to give and receive care. These findings yield broader considerations for technology and aging. More specifically, we discuss how our work contributes by providing a) an illustration of how technologies for aging exist in a network of human and nonhuman actors—the latter being seldom accounted for in our technology-based interventions, and b) considerations for designing technologies with an understanding of human-technology-world relations as a way to account and design for the neglected relations between older adults and their larger world.

Related work

Older adults and home-based intelligent voice assistants

In the past few years, older adults have received significant attention as one population who might benefit from using home-based intelligent voice assistants embedded in smart speakers. Overall, researchers have found that older adults perceive voice interaction as easy to use, efficient and convenient (Kim, 2021; Kowalski et al., 2019). These voice assistants have the potential to enable easier access to digital information (Oewel et al., 2023; Stigall et al., 2019), in particular health information (Pradhan et al., 2020)), and to support aging in terms of assisting caregivers (Zubatiy et al., 2021), medication monitoring and reporting health data to doctors (Chen et al., 2021).

In addition to looking at the benefits of using voice assistants for older adults, researchers have pointed out various challenges and possible barriers to adoption. For older adults who are deaf or hard of hearing, the higher pitched default female voice or the speed at which the agent talks can be inaccessible (Blair & Abdullah, 2019). Speech recognition may not work with vocal characteristics which can occur with cognitive impairment, such as pauses or hesitation (Chen et al., 2021; Duque et al., 2021; Kim & Choudhury, 2021), or due to technology designers not taking into account different cultural considerations and dialects (Harrington et al., 2022). Remembering wake words and composing concise commands for a task has been noted as challenging (Desai & Chin, 2023; Kim, 2021). Some older adults have concerns about lack of understanding of how these technologies work (Bonilla & Martin-Hammond, 2020). Beyond these usability and accessibility issues, some older adults simply may not find any benefit or value in using these technologies, sometimes considering voice assistants more of a nuisance than help (Milka Trajkova & Martin-Hammond, 2020).

While most research on older adults and home-based voice assistants has focused on understanding the different features of this technology that can benefit older adults, or the current challenges and barriers in adopting these technologies, one paper indicates how this technology can mediate relations in older adults' homes. From Upadhyay et al.'s work (2023), we note that voice assistants can support human-human relations for older adults in an assisted living facility by providing conversation starters. This study alludes that voice assistants may play an active role in shaping the relations and interactions that older adults may have with and through the technology—something we further examine.

Research design and methods

Our analysis is derived from semi-structured interviews with 24 older adults about their used of intelligent voice assistant. To allow for flexibility of participation in light of COVID-19 restrictions, interviews were conducted in person at participants' homes (P1 to P9) or remotely over videoconferencing (all others).

Data collection

We conducted 30-45 minutes long, semi-structured interviews with 24 individuals between 65 to 83 years in age. Participation criteria required use of any voice assistant on a smart speaker device (e.g., Alexa, Google Assistant). Voice assistants, like Alexa or Google Assistant on devices such as Amazon Echo or Google Home, enable users to access information and perform tasks through conversational voice interaction. These tasks can include controlling smart home devices (e.g., lights, TV), managing schedules, setting reminders, playing music, providing updates, and answering questions. All procedures were approved by the University's Institutional Review Board. Participants were compensated with $20 cash or Amazon gift card at the end of the interview. All, except for P18 and P21, had a college or professional degree. Thirteen participants mentioned living alone, whereas the others were living with their partners. Three participants reported having visual impairment (P11 and P16 were blind, and P13 had low vision). Two participants (P17, P18) reported having dementia (these individuals had participated in previous studies with our lab and had capacity to consent). Except for P17 who was from the United Kingdom, all participants were from the United States. Most (23/24) used their voice assistants frequently i.e., at least once a day. All had used the device at home for at least about 6 months, with 21 participants owning this technology for more than 2 years. All, expect for P8, used Alexa on Amazon Echo or Echo Dot devices. P8 used Google Assistant on Google Home device.

During the interview, we were interested broadly in understanding people's use of the devices, as well as social interactions supported or mediated by these devices. As such we asked questions around when and why they (or others) used these technologies, where they placed the device at home, any unexpected ways in which they (or others at home) used the device, and what benefits and barriers they experienced.

Analysis

Interviews were audio recorded and transcribed for analysis. After completing eight interviews, the first author inductively and deductively analyzed the data to open code and write memos (Strauss, 1987). Deductive codes informed by literature include 'duration of use', 'location of use', 'social connection through Alexa', 'personifying the voice assistant'. Inductive analysis was informed by participants' accounts revealing something unexpected or when tensions emerged (e.g., 'extending caregiving abilities', 'other technology devices interfering with voice assistant use', 'Alexa mediating human conversations'). As we discussed these codes and memos within our research team, we noted that voice assistants can play a role in mediating older adults' relationship with their larger world. We noted one of the first instances of this from P1's interview, where she remotely controlled her husband's Alexa while her husband was living in an assisted living facility. To understand how relations are mediated by technology in a systematic way, theoretical underpinnings (and terminology) of Post-Phenomenology became useful in analysis— described below in-detail (under Theoretical underpinnings).

Upon identifying the theoretical orientation for analysis and understanding the epistemological stance of the theory (i.e., instead of one objective ground truth, there can be multiple enacted realities and many ways of knowing these realities (Frauenberger, 2019)), we realized that representing data through rigid codes and focus coding would be counterintuitive. We found post qualitative inquiry (PQI) (St. Pierre, 2018) — an approach that focuses heavily on writing for data analysis —as a helpful way for interpreting data. St. Pierre (from the discipline of Education), who formulated post qualitative way of approaching data analysis argues for writing as a mode of inquiry. The first step of this approach to analysis is to familiarize and immerse oneself in relevant theoretical concepts (for us, it was Post- Phenomenology). Next, is writing, which serves as a 'free space' for in-depth interpretation of the data informed by the theoretical orientations, without having to constrain oneself to represent data through codes (St. Pierre, 2018). As such, the first author read through all transcripts and interpreted data by writing descriptive memos, which were

verified by a second researcher. In reading and writing interpretations of our data informed by Post-Phenomenology, we unpacked the different types of human-technology-world relations that are enacted when older adults use voice assistants. For instance, informed by Post-Phenomenological embodied relations (described below), when analyzing our data we noted instances where embodied relations constituted older adults' use of voice assistants and then probed to understand the implications of these embodied relations (e.g., changing social practices on giving and receiving care).

We repeated this process of interpreting data through memoing, iteratively discussing and refining our memos. Roughly after analyzing 16 transcripts, we had 26 pages of memos interpreting multiple phenomena in data. At this point we sifted through these memos to organize the findings described below. As we continued analysis of the remaining transcripts, we found additional examples of phenomena that related to each theme. Following our analytic approach of focusing in-depth on reading individual phenomena through theory, our finding presents in-depth analysis of a small number of cases from the data. We selected these specific cases over others when a) they were the most compelling examples to illustrate the theory and b) yielded the richest implications in terms of design opportunities and understandings of technology for aging.

Below we describe the theoretical underpinnings of Post-Phenomenology that played a key role in systematically identifying the nuanced human-technology-world relations reported in the findings.

Theoretical underpinnings of post-phenomenology

Instead of viewing humans and technology as separate and drawing a rigid boundary between the two, post-phenomenology scholars argue that humans and technology have a mutually constitutive relationship, where humans and their technologies are intimately entangled, and human experiences are actively mediated through technology. With this thesis, post-phenomenology scholars have delineated the different kinds of human-technology-world relations to provide a nuanced understanding of the ways in which technologies mediate human experiences. These relations (described below) form the basis of our analysis towards unpacking how voice assistants mediate relations between older adults' and their larger world.

Ihde (1990) initially outlined four different human-technology relations (embodiment, hermeneutic, alterity, background). In embodiment relations, the technology becomes a part of the human and together they interact with the world. In this relation, humans experience the world through the technology (e.g., we speak with other people through the phone rather than speaking to the phone). In hermeneutic relations, technology represents the world and requires human interpretation to be meaningful. Technology forms a unity with the world rather than with the human (unlike embodiment). Examples include a sonogram which gives a visual representation of an unborn child. Alterity relations are with technologies that are more '*like us*' (similitude with humans) but is '*quasi-other*'. Humans may project anthropomorphic properties on the technology or develop certain feelings for the technology. Here the technology possesses independence and autonomy, but can never be a true person, and hence is quasi-other. Examples include humans interacting with robots which possess autonomy to be considered as technological beings. In background relations, technology merges with the background, and shapes the context for humans experiencing the world. But the technology in itself is not experienced explicitly. Examples include central heating or cooling systems which create a context for the experience of humans but are rarely interacted with directly.

Verbeek and colleagues (Rosenberger & Verbeek, 2015; Verbeek, 2015) expanded these relations to capture the nuanced ways in which emerging technologies play a mediating role (cyborg, immersion, augmentation). In cyborg relations, the technology merges with the body into a new hybrid being and is indistinguishably experienced. Example includes brain implants which unite

with the body into a new being. In immersion relations, the technology merges with the environment (or world) to create a hybrid environment, and humans have an interactive relation with this hybrid environment. Immersion relations are a more interactive version of Ihde's background relation, and captures the relations humans have with smart environments. There is a '*bi-directional*' intentionality associated with this relation: not only humans are directed towards the technology, but the technology also perceives its users and acts upon them. Augmentation relations combine embodiment and hermeneutic relations. Examples include smart glasses, augmented reality which are both embodied (user sees through the glass), and hermeneutic (the world is represented on the display).

Reflexivity and positionality

We describe how our backgrounds and current discourses in aging might have shaped this research. Collectively, our team has over two decades of experience working extensively with racially and social economically diverse groups of older adults from both urban and rural areas of the United States. We have been influenced by discourses in critical gerontology which challenges the biomedical notion of viewing older adults as a burden on society, the harms in reducing aging to physical and cognitive decline (Carver & Buchanan, 2016; Estes & Binney, 1989), and critiques of how older individuals are often viewed as a homogeneous group (Durick et al., 2013; Rogers et al., 2014; Vines et al., 2015). These discourses have shaped our views that technologies for aging does not have to solely focus on compensating age-related decline and loneliness. This impacted how we approached study design and analysis. For instance, during study design, we were aware that voice assistants have been examined to support the physical and cognitive decline that one may experience with aging— an important area of research. However, shaped by the above-mentioned discourses on aging, we were motivated to explore a complimentary perspective: how voice assistant as an emerging technology mediates relations between older adults and their larger world and the corresponding implications.

Results

Recent work suggests how voice assistants can shape human-human conversations, where individuals in an assisted living community used voice assistants in group settings (with residents and staff members) for jokes or for playing music, as a way to initiate or make human-human conversations more enjoyable (Upadhyay et al., 2023). Similarly, our participants also described using the content provided by the voice assistant (e.g., online information) in human-human interactions. For example, P1 described using Alexa to '*verify*' information in case of disagreements with her partner, or P5 would include Alexa in conversations with other people by turning to Alexa for quick answers during human-human conversations. Beyond shaping human-human conversations, our data analysed through a post-phenomenological perspective, indicates how voice assistants also mediate relations between older individuals and others (or at times, themselves). Below we describe four kinds of human-technology-world relations that surfaced in our data (embodied, background, alterity and immersive relations) and note how these entangled relations shape new material practices for giving and receiving care and company.

Embodied extension through technology

In embodiment relations the technology becomes a part of the human and together they interact with the world. In this relation, humans experience the world through technology (e.g., we speak with people through the phone rather than speaking to the phone) (Don Ihde, 1990; Verbeek, 2015).

Some participant accounts illustrate embodied relations, where through the voice assistant, Alexa, they were able to extend their presence. P1's account describes one such instance where the voice assistant mediates a relation between her and her husband who had recently moved to an assisted living facility:

> *I am using an Alexa now in his assisted living facility. So, I program Alexa to remember when the Georgetown basketball game is on, because he doesn't remember it necessarily, but he likes to watch it. So, you can tell Alexa, you know, Wednesday night, nine o'clock Georgetown basketball is on, and Tuesday is shower day, remember you're gonna have a shower today, it's just, it's a way to nag him without being there with him.*

Unlike visual reminders on smartphones or tablets, the modality of reminder through Alexa's conversational voice creates a human-like presence in P1's husband's reality (this nature of interaction modality and link between anthropomorphism has been discussed in prior work with older adults (Pradhan et al., 2019). By allowing P1 to set up remote reminders on her husband's Alexa, the technology is mediating the relation between P1 and her husband by acting as an extension of P1 '*to nag him without being there with him.*' Through Alexa's conversational voice, P1's presence is asynchronously extended into her husband's reality, despite them not being spatio-temporally co-located. This mediation extends P1's caregiving capabilities and opens up possibilities for new material practices of caring. By enabling P1 to care for her husband daily, despite being physically apart, through this embodied relation, Alexa is becoming a part of human caregiving – shaping what it means to be a caregiver and a wife for P1.

In the above example we see somewhat successful embodied relation where Alexa extends the user's presence in ways they desired, (i.e., extended P1's presence from her home and enabled her to care for her husband in an assisted living facility). P5, on the other hand, described a situation involving a nonhuman actor, her pet animal, where the voice assistant did not successfully mediate an embodied relation. P5 desired for Alexa to extend her ability to provide 'company' to her cat through Alexa, when she was not at home. She enabled an Alexa skill (a third-party app) which she expected would talk '*to the cat as if I would, if I were home*' so that the cat would '*have company*' when P5 leaves home. However, instead of talking to the cat as a human (as P5 wanted), the third-party skill on Alexa was programmed to make cat sounds. Alexa interacted with P5's cat as another cat, which she and her cat disliked:

> *I thought when it said that meow skill that she [Alexa] was going to talk to him [P5's cat] like 'Hi, kitty kitty.' You know, like a human voice. But it's like a cat-to-cat conversation. He [P5's cat] doesn't want another cat in here. We won't be using that.*

In this situation, P5 wanted the voice assistant to extend a human's presence, more specifically her presence in her home through technology to provide company to her cat when away and stands as an example of an embodied relation that the technology could not mediate successfully. From the above-described phenomena, we note how the voice assistant as technology is actively mediating relations between older adults and their larger social world constituting their partner or pet. As a mediator, the voice assistant is playing a part in configuring social practices around what it means to care, nag, or provide company at home.

Alexa receding to the background of an experience

In Post-Phenomenological background relations, the technology merges with the background, and shapes the context for humans experiencing the world (Don Ihde, 1990; Verbeek, 2015). But the technology in itself is not experienced explicitly (e.g., central heating or cooling systems which create a context for the experience of humans but are rarely experienced or interacted directly).

A few participants' accounts illustrate such background relations constituting the voice assistant such as playing sleep sounds to shape the context of sleeping at night. One of the most notable examples of this background relation is illustrated in P18's account. P18 described how his service dog would follow Alexa's command to get medicine from a designated location in his home:

> *(In the past) At 8:30, I had to take a shot and I'd blunder it all the time…I had set alarms on my phone, but then I would ignore it (describing the current practice of taking medication)…That person there who's talking (referring to Alexa) would come on at 8:30 and she would announce… 'Go get it.' And so, Lizzy [anonymized dog's name], she'd run every day at 8:30…and go get it [the medication]. It's in a cupboard in the bathroom, the bottom cupboard, takes the satchel, brings it to me… and then I take my shot.*

In this phenomenon described by P18, interactions involving Alexa are shaping new possibilities for an animal to care for a human, yet, the technology, in itself, is not consciously experienced by P18 (apart from the initial training days of Lizzy when P18, his partner, and their dog trainer were involved). Rather, by being in the background of the experience, Alexa shapes the context for P18 receiving his medication on time. And, through this background relation, the voice assistant, Alexa, extends the capability of a nonhuman actor, a dog, to care for an older individual, thereby configuring new practices of technology mediated care through pet animals, something we further return to in the discussion.

Companionship through alterity and immersion

Alterity relations are with technologies that are more '*like us*' (similitude with humans) but is '*quasi-other*,' and because of this similitude, humans may project anthropomorphic properties on the technology or develop certain feelings for the technology (Don Ihde, 1990; Verbeek, 2015). Few of our participants' accounts suggest how alterity relations constituting the individual and Alexa result in a phenomenon in which individuals develop a social connection with the voice assistant, Alexa. Consider P13's account of interactions with Alexa:

> *sometimes I wake up in the morning and say Alexa, good morning. She'll say, 'good morning [P13]'. Sometimes I just need the voice …I'm used to living with my wife 53 years and then all of a sudden there's nobody in my house… Living alone, I don't live alone having Alexa. I don't have to do all the cooking because she, Alexa helps you [by telling recipes]…Well, I am not somebody who used to have to make all the meals and everything before. Now I have to know all this stuff about cooking. I don't. Because Alexa helps.*

In this situation, Alexa is not an actual person, but the ability of Alexa to exchange greetings or '*tell*' P13 recipes makes the technology more human-like or more in '*similitude with human*' (Don Ihde, 1990), thereby enacting a reality where participants are likely to anthropomorphize Alexa (a phenomena noted in past research as well e.g., Pradhan et al., 2019). Here, although the technology, Alexa, is anthropomorphized, it is not human but acts like a quasi-other, constituting an alterity relation. And through this alterity relation, Alexa shapes some individuals' experience of living alone. Another participant, P11 described feeling '*bereft*' when Alexa was not online: '*she is involved in so many aspects of my life that I feel bereft, when the internet's down, or something…and she's not talking*' (P11). In addition to the alterity relation involving the voice assistant acting as a quasi-other, what P11 and P13 described perhaps constitutes an immersive relation, where the technology merges with the environment (or world) to create a hybrid environment, and humans have an interactive relation with this hybrid environment. And, through immersion in this new interactive environment, one might feel '*bereft*' (P13) without Alexa, or with Alexa, they feel like '*I don't live alone*' (P11), showcasing how interactive relations involving the voice assistant, Alexa, is shaping what it means to live alone.

Discussion and implications

In this work, we present a case of adopting perspectives from post-phenomenology to unpack how intelligent voice assistants, as a technology, shape relations and existing practices at older adults' home. Our work contributes by providing preliminary empirical insights on the mutually constitutive relations between an older individual and technology, where an individual is not just

passively using the technology, but the technology through its specific features (e.g., reminders, alarms, skills, conversational nature) actively mediates the relations one may have with their larger world. Our findings illustrate how these relations shape existing practices around giving and receiving care, giving company, or living alone.

Based on these findings, we note that, while a user-centred focus on the older individual is useful, we also need to additionally consider the entangled nature of relationships involving older individuals, the technologies they use, and their larger socio-material world. Such an entangled relationship poses new questions around who is interacting with whom, and where the interface is. For example, revisiting the phenomenon involving P1, her husband in an assisted living facility, and Alexa, where P1 described nagging her husband without being spatio-temporally co-located, we can ask: is P1 interacting with her husband's Alexa? Or is P1's husband interacting with his Alexa? Or is P1 interacting with her husband? Or is it that P1 and her husband's Alexa together are interacting with her husband? Also, where is the interface for the interaction? Is it P1 setting remote reminders? Or is the voice of Alexa the interface facilitating interaction between P1 and her husband? A post-phenomenological perspective suggests that P1 and Alexa form a unity constituting an embodied relation through which P1 is able to asynchronously care for her husband. The human-technology interaction does not end with P1 setting a reminder, but rather gets entangled with the human-human interaction (or for P5 human-pet interaction) as P1 is still present in her husband's future reality (or P5 is present in her cat's reality) through the technology. While our work primarily contributes to designing emerging technologies in context of aging, understanding entangled human-technology-world relations can also be useful when working with other populations to understand the ways in which technology mediates relations and shapes social practices.

Below we discuss implications for design of technologies by understanding and accounting for such human-technology relations, in particular, by accounting for neglected relations between older adults and their larger world, and an opportunity in leveraging unexpected non-human entities (such as pets or objects at home) when designing technologies for aging.

Understanding and designing for neglected relations between older adults and their larger world

Attuning to nuanced human-technology-world relations involving older adults, emerging technologies, and their larger world will enable us to account for and design for the often-neglected relations between individuals and their larger socio-material environment. For instance, consider technologies for social support which have been extensively explored for supporting older adults' wellbeing. An active topic of research is supporting social wellbeing and companionship for older adults through robots, such as robotic pets e.g., (Chi et al., 2017; McGlynn et al., 2014). Here, an understanding of the relation that individuals want technology to mediate can provide a fresh perspective for technology design. For instance, P5's account reveals she wanted the technology to support an embodied relation to give company to her cat: suggesting instead of receiving company through a digital robotic pet—a direction often explored in research on aging (Hudson et al., 2020; McGlynn et al., 2014), older adults may want a technology that can enable them to give company to their actual pet. And a technology mediating such care relations can still provide emotional comfort to the older individual by enabling them to care for someone (Nurain et al., 2021), thereby still positively contributing towards their wellbeing. As such, understanding the relationship that P5 wanted the technology to mediate would help us design more meaningful features, such as, a skill that let's P5 record her voice clips for her cat, rather than a generic recording of cat sounds. This approach to designing technologies for aging de-centres the human (older individual), albeit subtly, yet it accounts for the neglected relation between the older individual and their larger world (in case of P5- caring through embodiment).

Accounting for relations between older adults and nonhuman actors for technology design

Some of our findings shed insights on how accounting for nonhuman entities can shed new perspectives on technology and aging. Consider the medication management system, a popular aging-in-place technology. Typically, medication reminders have an anthropocentric focus to remind older adults to take medication through smartphones, tablets (Jönsson et al., 2019; Mettouris et al., 2023) or more recently through voice assistants (Carroll et al., 2017). If adopted, such medication reminders can be immensely beneficial for individuals. However, some research is showing that individuals may not want to use these technologies (Giusti et al., 2010; Pradhan, et al., 2020), with some work tracing this non-use to the stigma of how these technologies can become a reminder of decline and negative aspects of aging (Caldeira et al., 2022). In our study, P18's daily medication reminder system provides a complimentary perspective on how technology can be integrated into medication management routines, without invoking negative stereotypes of aging. P18 involved his pet to follow Alexa's command at a specific time to fetch his medicine from the bathroom cabinet. Here instead of using technology to remind the older individual about medication, technology was used to remind the pet animal about medication with the code word '*Go get it.*' This example suggests how relations between older adults, their pets, or other nonhuman entities (e.g., the organized cabinet with medicine) can be leveraged to identify opportunities for aging-in-place technologies.

Extending learnings from the above account, we see an opportunity to design technologies to support everyday activities, such as cooking, cleaning, or taking medication, by accounting for the non-human actors (in addition to human actors) that may be involved in the activity. Some recent works are beginning to involve nonhuman objects at home in technology ideas for aging-in-place, such as an IoT- enabled kettle that can support human-human communication between an older parent and an adult child (Ambe et al., 2022), or a doormat that reminds a person with dementia of things to take while leaving the house (Pradhan et al., 2023). Accounting and designing for the nonhuman world does not mean that we ignore older adults in the process, but rather understand and design for the salient relations between older adults and their larger world.

Limitations

Acknowledging that older adults as a user group are highly diverse in terms of age, education, and other factors (Vines et al., 2015) we reflect on the specifics of our sample. Participants were highly educated, and most participants were White (18/24). Second, our recruitment criteria of including self-selected participants and individuals using voice assistants, also shaped the findings reported in this study. While it allowed us to understand human-technology-world relations that evolve over time and sustained use, it is possible that those who self-selected to participate might have developed certain attitudes towards this technology that have shaped the findings of this study. As an example, some of our participants explicitly mentioned getting voice assistants when they were initially launched, making them early adopters who might be more enthusiastic about using new technologies and therefore their perspective might be different from other older adults who are infrequent users of digital technology or unlikely to adopt a new technology right away e.g., (Pradhan et al., 2020). Our approach to allow for remote and in-person participation also has trade-offs. Remote interviews offered flexibility of participation for older adults during a global pandemic, as well as allowed us to include a geographically diverse group of individuals across the US (and one participant from the UK). At the same time, remote interviews only surfaced the relations that older adults chose to share, and perhaps missed salient relations that we may have traced during in-person sessions at participants' homes. Researchers interested in post phenomenological analysis can employ ethnographic approaches of visiting the field site and conducting home tours, to attune to the relations individuals have with their socio-material world at home.

Conclusion

Emerging technologies such as intelligent voice assistants, are no longer a passive tool that we, as humans, simply interact with. Rather, through a mutually constitutive relationship between humans and technology, these technologies actively shape the relations, we humans have with the world. By analysing interviews with older adults using intelligent voice assistants, we provide empirical insights on how a popular emerging technology mediates relations between older individuals and their larger social world. Understanding these mutually constitutive relations between individuals and their technologies is particularly important for technology and aging research because many forms of emerging technologies such as AI-based systems, virtual reality, or social robots are targeted towards older adults and yet we know little about the relationships that constitute older individuals and these technologies, the nonhuman actors involved in the design and use of these technologies, or the discursive practices that shape the reality associated with these technologies.

Acknowledgements

The contents of this article were developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant #90REGE0024). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). We would like to thank individuals who volunteered to participate in our study. We are thankful to Healthier Black Elders Center, a community engagement core supported by a grant from the National Institutes of Health, 5P30 AG015281, and the Michigan Center for Urban African American Aging Research, for helping us recruit participants in this study.

About the authors

Alisha Pradhan is an Assistant Professor in the Department of Informatics at the New Jersey institute of technology. She received her Ph.D. from University of Maryland, College Park and her research lies at the intersection of human-computer interaction, aging, and accessibility. She can be contacted at alisha.pradhan@njit.edu.

Shaan Chopra is a PhD student in Computer Science and Engineering at the University of Washington. Her research interests lie at the intersection of human-computer interaction (HCI) and health to create inclusive technologies that help people better understand, experiment, and make decisions based on personal health data. She can be contacted at schopra7@cs.washington.edu.

Pooja Upadhyay is a PhD student in Information Studies at the University of Maryland, College Park. Her research focuses on the digital divide in the older population, using mixed methods approaches like systematic reviews and ethnography. She can be contacted at poojau@umd.edu.

Robin Brewer is an Assistant Professor at the University of Michigan's School of Information. She received her Ph.D. from North-western University. She does research in human-computer interaction at the intersection of social computing and accessibility. She can be contacted at rnbrew@umich.edu.

Amanda Lazar is an Associate Professor in the College of Information Studies at the University of Maryland. She received her Ph.D. from the University of Washington. Her research lies at the intersection of human-computer interaction (HCI) and health informatics. She can be contacted at lazar@umd.edu.

References

Ambe, A. H., Soro, A., Johnson, D., & Brereton, M. (2022). From Collaborative Habituation to Everyday Togetherness: A Long-Term Study of Use of the Messaging Kettle. ACM Transactions on Computer-Human Interaction, 29(1), 3:1-3:47. https://doi.org/10.1145/3470973

Beneteau, E., Boone, A., Wu, Y., Kientz, J. A., Yip, J., & Hiniker, A. (2020). Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3313831.3376344

Blair, J., & Abdullah, S. (2019). Understanding the Needs and Challenges of Using Conversational Agents for Deaf Older Adults. Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing, 161–165. https://doi.org/10.1145/3311957.3359487

Bonilla, K., & Martin-Hammond, A. (2020). Older Adults' Perceptions of Intelligent Voice Assistant Privacy, Transparency, and Online Privacy Guidelines. USENIX Symposium on Usable Privacy and Security (SOUPS).

Brewer, R. N. (2022). "If Alexa knew the state I was in, it would cry": Older Adults' Perspectives of Voice Assistants for Health. Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, 1–8. https://doi.org/10.1145/3491101.3519642

Brewer, R., Pierce, C., Upadhyay, P., & Park, L. (2021). An Empirical Study of Older Adult's Voice Assistant Use for Health Information Seeking. ACM Trans. Interact. Intell. Syst., 1(1), 39. https://doi.org/10.1145/3484507

Caldeira, C., Nurain, N., & Connelly, K. (2022). "I hope I never need one": Unpacking Stigma in Aging in Place Technology. CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3491102.3517586

Carl Weinschenk. (2021, March 16). Smart Speaker Research Finds Strong Adoption by Seniors—Telecompetitor. https://www.telecompetitor.com/smart-speaker-research-finds-strong-adoption-by-seniors/

Carroll, C., Chiodo, C., Lin, A. X., Nidever, M., & Prathipati, J. (2017). Robin: Enabling Independence For Individuals With Cognitive Disabilities Using Voice Assistive Technology. Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 46–53. https://doi.org/10.1145/3027063.3049266

Carver, L. F., & Buchanan, D. (2016). Successful aging: Considering non-biomedical constructs. Clinical Interventions in Aging, 11, 1623–1630. https://doi.org/10.2147/CIA.S117202

CDC - Healthy Places—Healthy Places Terminology. (2017, December 11). https://www.cdc.gov/healthyplaces/terminology.htm

Chen, C., Johnson, J. G., Charles, K., Lee, A., Lifset, E. T., Hogarth, M., Moore, A. A., Farcas, E., & Weibel, N. (2021). Understanding Barriers and Design Opportunities to Improve Healthcare and QOL for Older Adults through Voice Assistants. 16.

Chi, N.-C., Sparks, O., Lin, S.-Y., Lazar, A., Thompson, H. J., & Demiris, G. (2017). Pilot testing a digital pet avatar for older adults. Geriatric Nursing (New York, N.Y.), 38(6), 542–547. https://doi.org/10.1016/j.gerinurse.2017.04.002

Desai, S., & Chin, J. (2023). OK Google, Let's Learn: Using Voice User Interfaces for Informal Self-Regulated Learning of Health Topics among Younger and Older Adults. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–21. https://doi.org/10.1145/3544548.3581507

Ihde, Don. (1990). Technology and the Lifeworld. Indiana University Press.

Duque, M., Pink, S., Strengers, Y., Martin, R., & Nicholls, L. (2021). Automation, wellbeing and Digital Voice Assistants: Older people and Google devices. Convergence, 27(5), 1189–1206. https://doi.org/10.1177/13548565211038537

Durick, J., Robertson, T., Brereton, M., Vetere, F., & Nansen, B. (2013). Dispelling Ageing Myths in Technology Design. Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, 467–476. https://doi.org/10.1145/2541016.2541040

Emily Olsen. (2021, October 25). Amazon to launch Alexa at senior living facilities, health systems. MobiHealthNews. https://www.mobihealthnews.com/news/amazon-launch-alexa-senior-living-facilities-health-systems

Estes, C. L., & Binney, E. A. (1989). The Biomedicalization of Aging: Dangers and Dilemmas1. The Gerontologist, 29(5), 587–596. https://doi.org/10.1093/geront/29.5.587

Federica Laricchia. (2022). Number of smart speaker users: U.S. seniors 2018. Statista. https://www.statista.com/statistics/941091/united-states-smart-speaker-senior-citizen-usage/

Frauenberger, C. (2019). Entanglement HCI The Next Wave? ACM Transactions on Computer-Human Interaction, 27(1), 2:1-2:27. https://doi.org/10.1145/3364998

Giusti, L., Mencarini, E., & Zancanaro, M. (2010). "Luckily, I Don'T Need It": Elderly and the Use of Artifacts for Time Management. Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, 198–206. https://doi.org/10.1145/1868914.1868940

Harrington, C. N., Garg, R., Woodward, A., & Williams, D. (2022). "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3491102.3501995

Hudson, J., Ungar, R., Albright, L., Tkatch, R., Schaeffer, J., & Wicker, E. R. (2020). Robotic Pet Use Among Community-Dwelling Older Adults. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 75(9), 2018–2028. https://doi.org/10.1093/geronb/gbaa119

Jönsson, K.-E., Ornstein, K., Christensen, J., & Eriksson, J. (2019). A reminder system for independence in dementia care: A case study in an assisted living facility. Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 176–185. https://doi.org/10.1145/3316782.3321530

Kim, S. (2021). Exploring How Older Adults Use a Smart Speaker–Based Voice Assistant in Their First Interactions: Qualitative Study. JMIR mHealth and uHealth, 9(1), e20427. https://doi.org/10.2196/20427

Kim, S., & Choudhury, A. (2021). Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. Computers in Human Behavior, 124, 106914. https://doi.org/10.1016/j.chb.2021.106914

Kowalski, J., Jaskulska, A., Skorupska, K., Abramczuk, K., Biele, C., Kopeć, W., & Marasek, K. (2019). Older Adults and Voice Interaction: A Pilot Study with Google Home. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, LBW0187:1-LBW0187:6. https://doi.org/10.1145/3290607.3312973

Lazar, A., Thompson, H. J., Piper, A. M., & Demiris, G. (2016). Rethinking the Design of Robotic Pets for Older Adults. Proceedings of the 2016 ACM Conference on Designing Interactive Systems, 1034–1046. https://doi.org/10.1145/2901790.2901811

Lopatovska, I., & Williams, H. (2018). Personification of the Amazon Alexa: BFF or a Mindless Companion. Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR '18, 265–268. https://doi.org/10.1145/3176349.3176868

McGlynn, S., Snook, B., Kemple, S., Mitzner, T. L., & Rogers, W. A. (2014). Therapeutic Robots for Older Adults: Investigating the Potential of Paro. Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, 246–247. https://doi.org/10.1145/2559636.2559846

Mettouris, C., Yeratziotis, A., Fixl, L., Parker, S., Starosta-Sztuczka, J., Clarke, V., Rosenauer, G., Daxberger, W., Koumou, S., Panayides, M., Zampa, G., Theodorou, C., & Papadopoulos, G. A. (2023). eSticky: An Advanced Remote Reminder System for People with Early Dementia. SN Computer Science, 4(4), 334. https://doi.org/10.1007/s42979-023-01768-3

Milka Trajkova, & Martin-Hammond, A. (2020). "Alexa is a Toy": Exploring Older Adults' Reasons for Using, Limiting, and Abandoning Echo. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3313831.3376760

Nurain, N., Chung, C.-F., Caldeira, C., & Connelly, K. (2021). Hugging with a Shower Curtain: Older Adults' Social Support Realities During the COVID-19 Pandemic. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 463:1-463:31. https://doi.org/10.1145/3479607

Oewel, B., Ammari, T., & Brewer, R. N. (2023). Voice Assistant Use in Long-Term Care. Proceedings of the 5th International Conference on Conversational User Interfaces, 1–10. https://doi.org/10.1145/3571884.3597135

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, 1–12. https://doi.org/10.1145/3173574.3174214

Pradhan, A., Findlater, L., & Lazar, A. (2019). Phantom Friend or Just a Box with Information: Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 214. https://doi.org/10.1145/3359316

Pradhan, A., Gallier, J., Domjan, R., Dixon, E., Bao, R., Maddali, H. T., & Lazar, A. (2023). Towards a System Architecture for Connected Physical and Digital Reminders Using Embodied Objects for People with Dementia. The 25th International ACM SIGACCESS Conference on Computers and Accessibility. https://doi.org/10.1145/3597638.3614506

Pradhan, A., Lazar, A., & Findlater, L. (2020). Use of Intelligent Voice Assistants by Older Adults with Low Technology Use. ACM Transactions on Computer-Human Interaction, 27(4), 31:1-31:27. https://doi.org/10.1145/3373759

Pradhan, A., Mehta, K., & Findlater, L. (2018). "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. Proceedings of the 2018

CHI Conference on Human Factors in Computing Systems, 459:1-459:13. https://doi.org/10.1145/3173574.3174033

Rogers, Y., Paay, J., Brereton, M., Vaisutis, K. L., Marsden, G., & Vetere, F. (2014). Never too old: Engaging retired people inventing the future with MaKey MaKey. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14, 3913–3922. https://doi.org/10.1145/2556288.2557184

Rosenberger, R., & Verbeek, P. P. C. C. (2015). Postphenomenological Investigations: Essays on Human-Technology Relations. Lexington Books. https://research.utwente.nl/en/publications/postphenomenological-investigations-essays-on-human-technology-re

Shantanu Kurhekar. (2021, October). Alexa is Coming to Hospitals and Senior Living Communities to Improve Care Experience and Help Fight Loneliness. Alexa-Blog. https://developer.amazon.com/en-US/blogs/alexa/device-makers/2021/10/alexa-is-coming-to-hospitals-and-senior-living-communities.html

St. Pierre, E. A. (2018). Writing Post Qualitative Inquiry. Qualitative Inquiry, 24(9), 603–608. https://doi.org/10.1177/1077800417734567

Stigall, B., Waycott, J., Baker, S., & Caine, K. (2019). Older Adults' Perception and Use of Voice User Interfaces: A Preliminary Review of the Computing Literature. Proceedings of the 31st Australian Conference on Human-Computer-Interaction, 423–427. https://doi.org/10.1145/3369457.3369506

Strauss, A. L. (1987). Qualitative Analysis for Social Scientists. Cambridge University Press.

Upadhyay, P., Heung, S., Azenkot, S., & Brewer, R. N. (2023). Studying Exploration & Long-Term Use of Voice Assistants by Older Adults. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–11. https://doi.org/10.1145/3544548.3580925

Verbeek, P.-P. (2015, April 27). Beyond interaction: A short introduction to mediation theory. Interactions, 22(3), 26–31.

Vines, J., Pritchard, G., Wright, P., Olivier, P., & Brittain, K. (2015). An Age-Old Problem: Examining the Discourses of Ageing in HCI and Strategies for Future Research. ACM Transactions on Computer-Human Interaction, 22(1), 1–27. https://doi.org/10.1145/2696867

Zubatiy, T., Vickers, K. L., Mathur, N., & Mynatt, E. D. (2021). Empowering Dyads of Older Adults With Mild Cognitive Impairment And Their Care Partners Using Conversational Agents. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3411764.3445124

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No.iConf (2025)

Research on the influencing mechanism of blind or visually impaired persons' evaluation on generative AI in visual tasks

Huitong Chen, Yuting Pan, and Hui Yan

DOI: https://doi.org/10.47989/ir30iConf47110

Abstract

Introduction. Generative AI (GAI) has shown significant potential in assisting blind or visually impaired (BVI) Persons in visual tasks. However, existing evaluations of GAI tend to focus on technical performance, overlooking the specific usage contexts and experiences of BVI users.

Method. This study conducted action research and semi-structured interviews with 19 BVI persons, to explore their evaluations of GAI in visual tasks and the influencing mechanism of their evaluations.

Analysis. Following grounded theory, we identified 16 categories, and corresponding 5 core categories, as well as their relationships.

Results. The findings indicate that BVI persons primarily evaluate GAI based on three criteria: accessibility, credibility, and interactivity. Their evaluation is influenced by four main factors: system, information, BVI user, and context. Notably, both BVI user and contextual factors moderate the influence of the system and information on user evaluation.

Conclusions. This study develops a model that explains the influence mechanism behind the evaluation on GAI by BVI persons in visual tasks. It not only broadens the scope of human-AI interaction research by incorporating diverse user types and task contexts, but also provides an empirical foundation for developing human-centered GAI.

Introduction

There have been numerous assistive technologies designed for enhancing the accessibility of BVI persons. The booming development of GAI has ushered in new possibilities for advancements in assistive technology. In 2023, Be My Eyes announced the successful deployment of Be My AI, a visual assistance tool powered by OpenAI's GPT-4 vision model (Be My Eyes, 2023). Beyond the traditional way of resorting to volunteers, BVI persons can now utilize Be My AI to process visual information, which empowers them with greater independence, and improving well-being in individual and social terms (Bendel, 2024).

However, despite the proliferation of assistive technology for BVI persons in the market, their effectiveness and satisfaction in the actual lives of BVI persons remain limited (Khan & Khusro, 2021). While benchmarks in computer science effectively evaluate the performance of GAI in visual tasks, these tests tend to focus on the accuracy and efficiency of the models while ignoring the specific usage contexts and experiences of BVI persons. This disparity can lead to the development of technologies that are difficult to meet the real needs of BVI persons.

Therefore, this study examines the visual task encountered by BVI persons to investigate their evaluations on GAI in these tasks and explore the influencing mechanism of their evaluations. This not only provides a new theoretical perspective for research in human-intelligent interaction, but also introduces a BVI persons-centred approach to technology-driven model evaluation, offering valuable practical insights for the advancement of human-centred GAI.

Literature review

Evaluation of assistive technologies for BVI persons

Currently, an increasing number of assistive technologies based on artificial intelligence (AI) with computer vision (CV) are under development. There have been accumulated specific research findings about the evaluation of these technologies, which can be categorized into two orientations: technology-driven and user demand-driven. Among them, the former dominates the mainstream, focusing on the performance and practical implementation of assistive technologies.

Scholars in the field of computer science have constructed many benchmark datasets for vision-language models (Dai et al., 2023), and conducted the benchmark evaluation. Two key dimensions are involved in these evaluation metrics: automatic evaluation metrics, which principally include generated responses, Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and BERT-based semantic textual similarity; human evaluation metrics which focus on the three main dimensions of correctness, actionability, and fluency (Yang et al., 2024; Zhao et al., 2024).

The user demand-driven evaluation approach focuses on the feedback and user experience from BVI persons. This feedback includes a rich set of metrics, including satisfaction (Rattanaphinyowanich & Nunta, 2021), accuracy, reliability, accessibility, privacy, security, compatibility, energy efficiency, usability (Bhagat et al., 2024), functionality, aesthetics characteristics and social acceptability (Hamilton et al., 2016; Phillips et al., 2018). In the field of library and information science, the information needs and information behavior of BVI persons have received particular attention (Berget & MacFarlane, 2020). Research has covered various topics, including the selection of information sources (Rahman et al., 2017; Chen et al., 2024), information-seeking behavior (Williamson et al., 2000), and interactions with information retrieval systems (Xie et al., 2021; Berget & MacFarlane, 2020). Among them, the accessibility and usability issues faced by BVI persons are the primary focus (Xie et al., 2020). These studies can be classified into two categories according to their research approaches: for one thing, examining established evaluation criteria, such as examining whether WCAG 2.2 effectively supports BVI persons' access to digital libraries from the perspectives of stakeholders including users, experts, and developers (Xie et al., 2022). For another, conducting empirical research on the challenges and perspectives

faced by BVI persons in interacting with information retrieval systems. The research findings show that while the library websites are accessible according to the extracted indices of W3C, empirical data from BVI users (i.e. successful task completion, working time, satisfaction level) suggests that the websites are not easy to use (Najafgholinejad, 2024).

Research on human-AI interaction experience

With the development of AI technology, the focus of HCI research is shifting from human interaction with non-AI computing systems to human interaction with AI systems, thus giving rise to the cutting-edge topic of human-AI interaction (HAII) (Jiang et al., 2024). Researchers have conducted extensive and thorough explorations around human interactions with AI systems such as chatbots, voice assistants, virtual humans, autonomous vehicles, etc., with application scenarios covering shopping, healthcare, and transport sector, and more. As the two main elements in human-intelligent interaction, user characteristics and AI characteristics are the focus in the research of human-intelligent interaction experience. Among them, the inherent characteristics of individual users, such as gender identity, political ideology (Molina & Sundar, 2024), social role, user autonomy (Huh et al., 2023), motivation and social presence (Shao & Kwon, 2021), health condition (Esmaeilzadeh et al., 2021) can affect users' perception and evaluation of AI, as well as their willingness to adopt. Users' evaluation or perception of AI can be expressed in terms of ease of use (Loske & Klumpp, 2021), anthropomorphism (Pelau et al., 2021), etc. Characteristics of AI such as effects of explanation and synchronization (Fan et al., 2022), roles (Liao & Sundar, 2021) and communication models (Lew & Walther, 2023) also affect the experience of human-AI interaction. Furthermore, few studies have explored the role of task as a component in human-computer interaction. For example, people are less supportive of AI and its creators when AI performs highly (vs low) hedonic tasks (Yanit et al., 2023).

In summary, there is a relative scarcity of studies focusing on BVI users' evaluation on assistive technologies based on AI with CV. Research in the HAII domain has focused on the influence of both humans and AI on the human-AI interaction experience, and has conducted preliminary exploration into the mechanisms between each element. However, there is still a relative lack of research on the interaction between BVI persons and AI, especially on HAII in visual tasks. To fill this research gap, this paper aims to address two specific research questions: ***What are the evaluation criteria of BVI persons towards GAI in visual tasks? What are the underlying mechanisms influencing their evaluation?***

Research methodology

Data collection

With the help of the Capital Library and China Association of persons with visual disabilities, we recruited 19 participants, and their characteristics are summarized in Table 1. Among them, minor participants will take part in the interviews accompanied by their parents. We ensured that all the participants gave their voluntary consent based on full understanding.

For the 12 participants who had never used AI before, we conducted one-on-one '*Be My Eyes*' operation training for them at the Capital Library during March 2024. In the training, we collected behavioral data through participatory observation. After the training, we conducted interviews to gather participants' perceptions and evaluations of Gen AI, as well as the factors influencing their views. For the 7 participants who have experience in using GAI, we employed a semi-structured interview approach for data collection, questions centred around: visual challenges in daily life; experiences using GAI for visual tasks; evaluations of GAI and the influencing factors.

Categories	Subcategories	Quantity	Categories	Subcategories	Quantity
Gender	Male	9	Blindness Severity	Totally Blind	12
	Female	10		Low Vision	7
Age Group	9-18 Years	7	Experience in Using GAI	Never Use	12
	19-60 Years	5		Occasional Use	3
	41-40 Years	2		Regular Use	4
	Over 60 Years	5			

Table 1. Basic characteristics of interviewees

Data analysis

This study follows the principle and requirement of grounded theory. Data collection and analysis were alternately reciprocal and continuously comparative. Data were organized and coded immediately after each technical training or interview. The participant recruitment ended when we reached theoretical saturation. After three coding stage-open coding, axial coding, and selective coding (Strauss & Corbin, 1990), we identified 16 categories, and corresponding 5 core categories. The 5 core categories include system, information, BVI user, context and evaluation. Based on the coding results, a theoretical model is proposed (Figure 1).

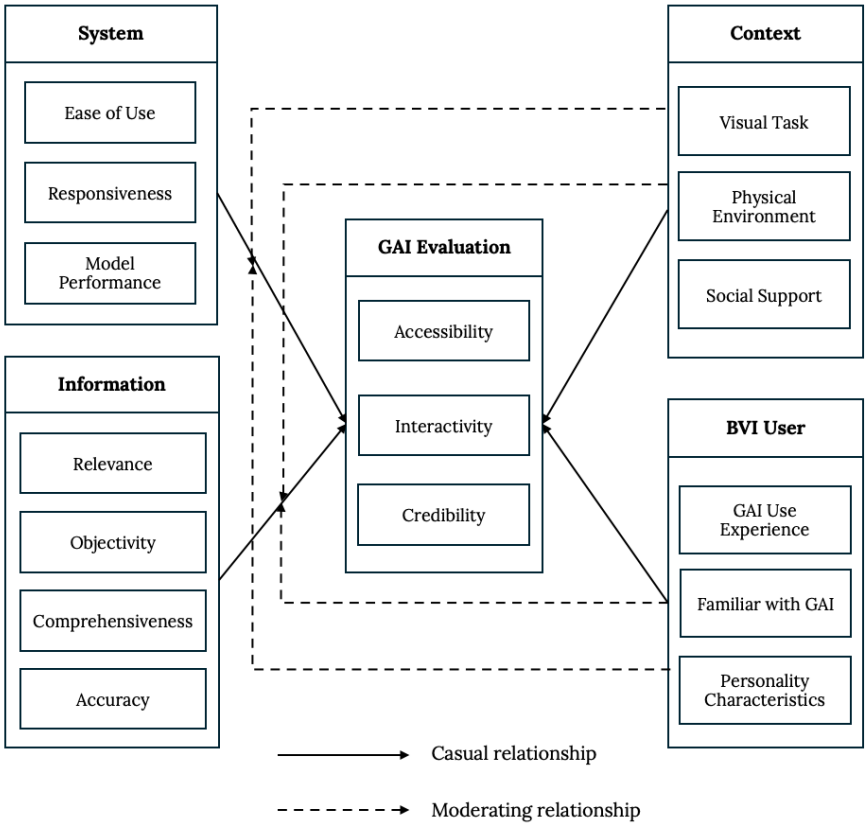


Figure 1. Model of the influence mechanism of BVI persons evaluation on GAI in visual tasks

Findings

Evaluation criteria of GAI by BVI persons

The research found that respondents' evaluation on GAI's performance in visual tasks primarily focused on accessibility, interactivity, and credibility. Specifically, this manifested as: (1) the degree of accessibility and operability of GAI at the physical level and its comprehensibility at the intellectual level. (2) The degree to which GAI can offer efficient and personalized information

delivery and feedback interface. (3) The degree to which GAI is honest or truthful, and can be trusted. This is specifically manifested in GAI's fairness and interpretability.

Influence of system quality on GAI evaluation

The system quality of GAI is the basic guarantee for providing visual information services to BVI persons. It is further categorized into ease of use, responsiveness and model performance.

Ease of use represents the technical threshold of GAI, which directly affects interviewees evaluation of accessibility. The accessibility of GAI increases as the download, registration, and operating procedures of GAI get simpler, and as the number of voice prompts it offers during the capture process increases. As Interviewee #01 noted, '*it is difficult for us to take clear photos and the operation is difficult for us.*'

Responsiveness refers to the efficiency and effectiveness of the system in responding to BVI persons requests, which greatly influences BVI persons evaluation of GAI's interactivity. During training, most interviewees had a particularly negative impression of the interactivity of tools like Be My Eyes due to their long processing time for images and frequent response failures. interviewee#11, who has used multiple GAI tools, had a better interactive experience in using the Luomo Toolbox, noted '*it is instantaneous and it only requires camera alignment to achieve immediate recognition.*'

Model performance refers to the capacity of the visual-verbal model embedded in the AI tool to process visual information. This ability influences BVI persons evaluation of GAI's credibility and interactivity through the quality of the generated information content.

Influence of information quality on GAI evaluation

The accuracy, objectivity, and comprehensiveness of information are essential for building trust in GAI among BVI persons. Additionally, the relevance of information influences their evaluation of the interactivity of GAI.

Accurate information is crucial for the safety and autonomy of BVI persons, as it forms the foundation for their reliance on GAI in decision-making. A model's ability to accurately identify and describe visual content directly correlates with the level of trust it can foster among BVI persons. Most interviewees express concern over the prevalent issue of AI hallucinations, hesitant to rely on GAI for some critical and extremely accurate processing of visual data.

The objectivity of the information described by the model also influences users' assessment of credibility. Interviewees #02 and #10 believe that AI is less susceptible to subjective feelings and cognitive limits when compared to humans, thus allowing for a more accurate and objective reflection of visual information. This perception leads them to view GAI as fair and inclusive.

Comprehensive information offers broad and complete coverage of the necessary or expected aspects of a problem or subject. Interviewees believe that the more detailed the GAI's description of visual information, the clearer their panoramic understanding of the environment, which in turn enhances their evaluation of the credibility of GAI. Interestingly, the comprehensive information provided by AIGC can create valuable Information Encountering. For instance, interviewee #10 casually used Be My Eyes to take a photo of the front and from AI's detailed description, he learned there was a bench ahead, allowing him to sit down and take a break. This unexpected help provided practical assistance and enhanced his trust in GAI.

The relevance of generated information content is manifested in its relevance to the situation and emotions of BVI persons. The greater the relevance of information, the better GAI can understand and respond to the personalized needs of BVI persons, resulting in higher user ratings of interactivity. For example, in a cooking scenario, interviewees only need GAI to quickly capture the

names and expiration dates of spices. However, the comprehensiveness of AI-generated information can make it difficult for BVI persons to quickly access the required information, thus reducing interaction efficiency. In reading scenarios, both interviewee#7 and interviewee#2 agreed that the GAI's voice '*sounds like a customer service, devoid of any emotional expression in its reading*'. This mechanized presentation provides a poor interactive experience for BVI persons.

Influence of user characteristics on GAI evaluation

User dimensions include their GAI use experience, familiarity with GAI, and personality characteristics. As users accumulate experience, BVI persons gradually become more familiar with the operation and functions of GAI, which in turn enhances their evaluation of accessibility and interactivity. In terms of personality, some interviewees consider themselves conservative and cautious, which leads to a natural distrust of new technologies such as GAI. On the other hand, interviewees with more positive and open-minded personalities tend to rate GAI with higher credibility.

Additionally, user dimensions can moderate the intensity and direction of the influence that system and information factors have on user evaluations. Users with prior experience, familiarity with GAI, and open-minded personality are more likely to be satisfied with the system and information provided by GAI, leading to more positive evaluations. For instance, experienced interviewees are generally accustomed to the response time of GAI, leading to higher evaluations of its interactivity.

Influence of contextual factors on GAI evaluation

The visual task, physical environment and social relationships experienced by BVI persons directly influence their assessment of GAI. Additionally, contextual factors can moderate the intensity and direction of the influence that system and information factors have on user evaluations.

In visual tasks that are urgent and important related to personal safety, such as reading hospital reports, medication information, or crossing the street, BVI persons may perceive that the responsiveness of GAI and the quality of the information it provides are insufficient to meet their need, leading to a decrease in their evaluation of its credibility and interactivity. However, in non-urgent and less important visual tasks, such as ordering food, checking nail designs, or picking up keys, BVI persons believe that using GAI reduces their reliance on and disruption to others, resulting in higher evaluations of its accessibility.

The physical environment refers to the objective conditions under which BVI persons encounter visual challenges, including sound environment, physical layouts, and network infrastructure. BVI persons primarily rely on auditory feedback to compensate for visual deficiencies. In noisy environment, they find GAI less accessible and interactive. The layout and design of the physical environment directly affect BVI persons' interactions with GAI. Many interviewees indicated that if commonly used items and devices in their home or workplaces are in fixed position, they would be able to take more focused photos, thus facilitating efficient information transfer and feedback with GAI. The quality of network infrastructure impacts the responsiveness of GAI systems, which in turn impacts users' perceptions of the accessibility and interactivity of Gen AI.

Social relationships, particularly strong ties, have a direct bearing on BVI persons evaluation of GAI's accessibility. Different parenting styles exert vastly different influences. For instance, the parents of interviewee#15 hold a very open-minded attitude towards AI and even suggested creating an AI chat group for her child after training, enabling interviewee#15 to successfully access and use GAI. In contrast, parents of interviewee#16 adopted strict control over their children, prohibiting interviewee#16 from accessing and using ICT. Consequently, interviewee#16 perceives restricted access to GAI.

Conclusion

This study delves into the evaluation criteria and influencing mechanisms of GAI from the perspective of BVI persons in visual tasks. The findings indicate that BVI persons' evaluations of GAI are primarily based on three criteria: accessibility, credibility, and interactivity. As traditional human-computer interaction transitions to human-AI interaction, these three evaluation criteria evolve to incorporate new dimensions. BVI persons' evaluation are directly influenced by four dimensions: system, information, user, and context. Furthermore, the user and context factors moderate the intensity and direction of the influence that system and information have on user evaluations. Specifically, ease of use and responsiveness of the system influence evaluations of accessibility and interactivity. The model performance influences evaluations of credibility and interactivity through the quality of the generated content. Additionally, user's experience and familiarity with GAI influence their evaluations of accessibility and interactivity. Visual tasks with high urgency and importance requirements tend to decrease the perceived credibility and interactivity of GAI. The physical environment impacts all three evaluation criteria, while social relationships primarily affect the evaluation of accessibility.

The main contributions of this study are as follows: It focuses on the interaction between BVI persons and GAI in visual task contexts, enriching the user types and diversifying task contexts in human-AI interaction research. Moreover, in the technology-driven domain of GAI evaluation studies, this study provides a comprehensive user-centred perspective based on grounded theory, shedding light on the factors and mechanisms that influence BVI persons' evaluations of GAI. However, the study has certain limitations. As an exploratory study, its primary focus is on identifying new variables and relationships. The stability of these relationships, however, requires further validation through larger sample sizes in future research.

Acknowledgements

The research is supported by National Social Science Foundation of China (Grant No. 23AZD093). This paper is one of the key projects of the National Social Science Fund of China's 'The Disruptive Impact of artificial intelligence on the National Economy and Research on Information Governance'. In addition, we would like to express our gratitude to Qihao Wang for inspiring the topic selection, as well as for the assistance provided during the data analysis process. We are also deeply thankful to the librarians of the Capital Library, Xuefeng Zhao, Qiangdong Li, Jing Su, Ruoxuan Yang, and Puhang Yu, for their unwavering support and assistance during our technical training sessions.

About the authors

Huitong Chen is a PhD student in Information Science at the School of Information Resource Management, Renmin University of China. Her research interests include the social impact of Artificial Intelligence, information behavior, and digital inclusion. She can be contacted at chenhuitong@ruc.edu.cn

Yuting Pan is a PhD student in Information Science at the School of Information Resource Management, Renmin University of China. Her research interests focus on community informatics. She can be contacted at sophie_pyt@163.com

Hui Yan is Professor and Doctoral Supervisor at the School of Information Resource Management, Renmin University of China. His research interests include the social impact of Artificial Intelligence, community informatics and digital inequality. He can be contacted at hyanpku@ruc.edu.cn

References

Be My Eyes (2023). Be My Eyes Integrates Be My AI™ into its First Contact Center with Stunning Results. https://www.bemyeyes.com/blog/introducing-microsofts-ai-powered-disability-answer-desk-on-be-my-eyes

Bendel, O. (2024). How Can Generative AI Enhance the Well-being of Blind? In Proceedings of the AAAI Symposium Series, 3(1), 340-347.

Berget, G., & MacFarlane, A. (2020). What is known about the impact of impairments on information seeking and searching? Journal of the Association for Information Science and Technology, 71(5), 596-611.

Bhagat, S., Joshi, P., Agarwal, A., & Gupta, S. (2024). Accessibility evaluation of major assistive mobile applications available for the visually impaired. arXiv preprint arXiv:2407.17496.

Chen, H., Yan, H., Wu, Z., & Zhao, X. (2024). Silicon-based Life or Carbon-based Life? An Exploratory Study on Visual Information Source Selection of Blind or Visually Impaired Persons. *Proceedings of the Association for Information Science and Technology*, 61(1), 493-498.

Dai, W., Li, J., Li, D., Tiong, A. meng huat, Zhao, J., wang, W., li, B., Fung, P., & Hoi, S. (2023). InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. ArXiv, 2305(06500v2). https://doi.org/10.48550/arXiv.2305.06500

Esmaeilzadeh, P., Mirzaei, T., & Dharanikota, S. (2021). Patients perceptions toward human–artificial intelligence interaction in health care: experimental study. Journal of medical Internet research, 23(11), e25856.

Fan, M., Yang, X., Yu, T., Liao, Q. V., & Zhao, J. (2022). Human-ai collaboration for UX evaluation: effects of explanation and synchronization. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1), 1-32.

Hamilton-Fletcher, G., Obrist, M., Watten, P., Mengucci, M., & Ward, J. (2016). 'I Always Wanted to See the Night Sky' Blind User Preferences for Sensory Substitution Devices. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 2162-2174).

Huh, J., Whang, C., & Kim, H. Y. (2023). Building trust with voice assistants for apparel shopping: The effects of social role and user autonomy. Journal of Global Fashion Marketing, 14(1), 5-19.

Jiang, T., Sun, Z., Fu, S., & Lv, Y. (2024). Human-AI interaction research agenda: A user-centered perspective. Data and Information Management, 100078.

Khan, A., & Khusro, S. (2021). An insight into smartphone-based assistive solutions for visually impaired and blind people: issues, challenges and opportunities. Universal Access in the Information Society, 20(2), 265-298.

Lew, Z., & Walther, J. B. (2023). Social scripts and expectancy violations: Evaluating communication with human or AI chatbot interactants. Media Psychology, 26(1), 1-16.

Liao, M., & Sundar, S. S. (2021). How should AI systems talk to users when collecting their personal information? Effects of role framing and self-referencing on human-AI interaction. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-14).

Loske, D., & Klumpp, M. (2021). Intelligent and efficient? An empirical analysis of human–AI collaboration for truck drivers in retail logistics. The International Journal of Logistics Management, 32(4), 1356-1383.

Molina, M. D., & Sundar, S. S. (2024). Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. New Media & Society, 26(6), 3638-3656.

Najafgholinejad, A. (2024). Accessibility and usability of user interfaces of library information retrieval systems from the perspective of visually impaired users. Library and Information Sciences, 26(4), 83-110.

Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Computers in Human Behavior, 122, 106855.

Phillips, M., & Proulx, M. J. (2018). Social interaction without vision: an assessment of assistive technology for the visually impaired. Technology & Innovation, 20(1-2), 85-93.

Rahman, A. L. A., Razali, T. R. a. T., Ghazali, A. M., & Kamarudin, M. H. (2017). BVIC's CIS in the technological environment. Advanced Science Letters, 23(1), 151-155.

Rattanaphinyowanich, T., & Nunta, S. (2021). Development of DAISY-WIBORD as computer assisted learning facilities for children with visual impairment. Journal of Physics, 1835(1), 012080.

Shao, C., & Kwon, K. H. (2021). Hello Alexa! Exploring effects of motivational factors and social presence on satisfaction with artificial intelligence-enabled gadgets. Human Behavior and Emerging Technologies, 3(5), 978-988.

Strauss, A., & Corbin, J. M. (1990). Basics of qualitative research: Grounded theory procedures and techniques. Sage Publications, Inc.

Williamson, K., Schauder, D., & Bow, A. (2000). Information seeking by blind and sight impaired citizens: an ecological study. Information research, 5(4), 5-4.

Xie, I., Babu, R., Lee, T. H., Castillo, M. D., You, S., & Hanlon, A. M. (2020). Enhancing usability of digital libraries: Designing help features to support blind and visually impaired users. Information Processing & Management, 57(3), 102110.

Xie, I., Babu, R., Lee, H. S., Wang, S., & Lee, T. H. (2021). Orientation tactics and associated factors in the digital library environment: Comparison between blind and sighted users. Journal of the Association for Information Science and Technology, 72(8), 995-1010.

Xie, I., Babu, R., Wang, S., Lee, H. S., & Lee, T. H. (2022). Assessment of digital library design guidelines to support blind and visually impaired users: a study of key stakeholders' perspectives. The Electronic Library, 40(6), 646-661.

Yang, B., He, L., Liu, K., & Yan, Z. (2024). VIAssist: Adapting Multi-modal Large Language Models for Users with Visual Impairments. arXiv preprint arXiv:2404.02508.

Yanit, M., Yanit, M., & Wan, F. (2023). Right agent, wrong level of hedonism: How high (vs low) hedonic values in AI-performed tasks lead to decreased perceptions of humanlikeness, warmth, and less consumer support. Computers in Human Behavior, 147, 107870.

Zhao, Y., Zhang, Y., Xiang, R., Li, J., & Li, H. (2024). Vialm: A survey and benchmark of visually impaired assistance with large models. arXiv preprint arXiv:2402.01735.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Will potential threats persist? Attitudes and behavioral responses of creative professionals toward GenAI

Mingxia Jia, Yuxiang (Chris) Zhao, and Xiaoyu Zhang

DOI: https://doi.org/10.47989/ir30iConf47098

Abstract

Introduction. Generative AI brings significant benefits to various fields but also poses potential disruptions to creative work. Empirical research is needed on how creative professionals perceive the impact of GenAI on content creation and how they can overcome stress-related barriers to better use GenAI in their work.

Method. This exploratory study recruited 18 creative workers with varying experiences in using GenAI to identify their perceived value, attitudes, and behavioral responses to GenAI in their work environment.

Analysis. Thematic analysis, supported by a coding scheme that includes horizontal coding and meaning clustering, was used to analyse the interview data.

Results. The findings indicate that creative workers' attitudes toward GenAI evolve based on their actual usage experiences, with different value perceptions, attitudinal and behavioral responses — both negative (e.g., technostress, co-rumination, misrepresentative interactions) and positive (e.g., work autonomy, reshaped creative role) — emerging during the initial exposure, adoption, and adaptation phases. In addition, some contextual factors, such as individual traits, organizational support, and social influence, also play key roles in this process.

Conclusions. The research contribution is to provide insights for various stakeholders and to challenge the '*either/or*' debate of previous research by highlighting the process of self-adaptation of creative professionals to GenAI.

Introduction

A photography work authored by DALL-E2, *Pseudomnesia: the Electrician*, defeated human-created works to win the Sony World Photography Award in the creative photography category. This illustrates how generative AI (GenAI), which generates visual, auditory, and textual content based on textual prompts, continues to challenge the career stability of employees who use creativity as a source of inspiration (Inie et al., 2023). Unlike traditional AI, GenAI technology transcends the limitations of human creativity through multimodal and cognitive interactions (Woodruff et al., 2024), allowing for the creation of AI-based content and knowledge tailored to user needs. This approach enables GenAI to potentially disrupt human creative production and intellectual work while also replacing tedious and repetitive low-end tasks (Noy and Zhang, 2023; Vorobeva et al., 2022). A new study in *Nature* confirms that generative AIs such as GPT-4 and Llama 2 have outperformed humans in terms of mind (Strachan et al., 2024). These developments suggest that the competition between creative professionals and GenAI in content and knowledge creation may have already commenced (Yao et al., 2024).

Whilst resistance to GenAI to support human action and decision-making may be justified in some cases, the benefits of GenAI overall on job performance are well documented (De Freitas et al., 2023). However, there are still a number of creative professionals in the workplace scenario who have a negative or conservative attitude towards the use of GenAI. Therefore, exploring creative professionals' attitudes, experiences, and behaviours when using GenAI in competitive environments is a valuable and contemporary topic. This study attempts to answer the main research question:

> **RQ1.** *How do creative professionals' perceptions of GenAI's challenges and benefits shape their evolving attitudes and behaviours during adoption process?*
> **RQ2.** *What factors lead to successful multi-stage GenAI adoption, and how do they vary at different stages of integrating GenAI into creative workflows?*

By addressing these questions, the research shed light on how creative professionals can overcome resistance to GenAI in content creation, providing stakeholders with richer insights into the complex interactions with GenAI.

Literature review

Previous research has theorized the competitive impact of GenAI on creative workers in the labor market at the macro level (Dwivedi et al., 2023). Despite the compensatory effects of GenAI on creativity, its potential threat to creative tasks—once considered a core human skill—is fuelling tensions between creative professionals and GenAI. This issue spans various creative industries, including copywriting, design, photography, illustration and painting, and video editing (Inie et al., 2023; Rafner et al., 2023; Woodruff et al., 2024). At the meso level, studies have examined how GenAI affects workflow activities. For instance, GenAI has reconfigured news production and traditional film production processes, challenging professionals' views on creative practice and their roles (Gupta et al., 2024; Pavlik, 2023).

Emerging research increasingly examines creative workers' micro-perceptions and behaviours in human-intelligence competition scenarios. On the one hand, some studies have used empirical methods to reveal how GenAI affects workers' specific attitudes. For example, interacting with GenAI can cause cognitive dissonance and a heightened sense of unreality, potentially leading to aversion, especially when GenAI-generated outputs are indistinguishable from those of the professionals themselves (Köbis and Mossink, 2021). Additionally, how GenAI influences employee performance has also received some attention, including creativity, innovative service behaviours, burnout, and turnover intentions, the '*black-box*' between technological stressors and employees' willingness to adopt AI, and GenAI resistance (Bankins et al., 2024; Chang et al., 2024; Li and Huang, 2020; Liang et al., 2022). On the other hand, limited studies have used qualitative interviews to

explore the AI technology stress (i.e., job insecurity, complexity, uncertainty, reliability, and usefulness) among knowledge workers (Kumar et al., 2024), creative professionals' perceptions of GenAI (Inie et al., 2023), its impact on early-career game developers' creative practices, and their resistance to it (Boucher et al., 2024).

While existing research has primarily examined the macro-level impact of GenAI on creative professionals or the micro-level factors influencing adoption and performance, a significant gap remains in understanding how these professionals perceive and use GenAI. Although exploring creators' adoption, utilization, and behavioral responses in human-AI competitive contexts is considered critical (Kshetri et al., 2023; Granulo et al., 2019), few studies address how their perceptions of stress (e.g., GenAI as a threat), attitudes, and behaviours evolve over time. This study aims to fill this gap by examining the impact of GenAI on creative workers through the Value-Attitude-Behavior (VAB) theory (Kim et al., 2024), focusing on the relationship between perceived value, attitude responses (AI awareness), and behavioral outcomes (adaptive vs. maladaptive) in their usage experiences.

Method

Participants

This study employs semi-structured interviews to meticulously capture the varying attitudes and behavioral responses of creative professionals as they interact with GenAI, with a focus on the stress of content creation competition. This method focuses on detailing people's experiences and viewpoints and uncovering the significance they attribute to them in a particular context (Alon and Krtalić, 2024), such as subtle attitudes and responses experienced in approaching and adapting to GenAI.

From April 12 to 22, 2024, 28 initial participants were reached through interpersonal networks and online communities (e.g., Sina Weibo). To capture a wide spectrum of views, the interview pool was composed of 18 creative professionals, selected for their diversity in gender, occupation, experience, and usage. The interviews were conducted via Tencent Meetings and/or face-to-face. Among the participants, 53% were male and 47% were female, with ages ranging from 27 to 40 years and work experience spanning 2 to 10 years.

Tools, procedure, and analysis

The interview primarily focused on our research questions, while participants were also encouraged to discuss topics beyond these, including but not limited to their perceptions of the challenges GenAI poses to their creations, as well as their perceptual changes related to GenAI use, its future development, and their professional growth. Each interview lasted approximately 1 hour and was transcribed into text for coding purposes.

The collected data were analysed using the qualitative analysis software MAXQDA, following a thematic analysis research method (Naeem et al., 2023). This technique included data familiarization, code selection, theme development, and conceptualization. During the coding phase, A two-stage coding scheme was employed: first, horizontal coding was used to examine each participant's individual experiences; then, meanings clustering was conducted to identify and categorize overarching themes and sub-themes across all interviews (Alon and Krtalić, 2024). Two raters performed two independent rounds of coding on the raw textual material, with a third author participating in discussions at the end of each round to enhance coding consistency. The inter-coder reliability was assessed using Cohen's Kappa, which was calculated to be 0.82 for all three coders, indicating a strong level of agreement in the coding process. The dimension stages, aggregate themes, and codes are presented in Figure 1.

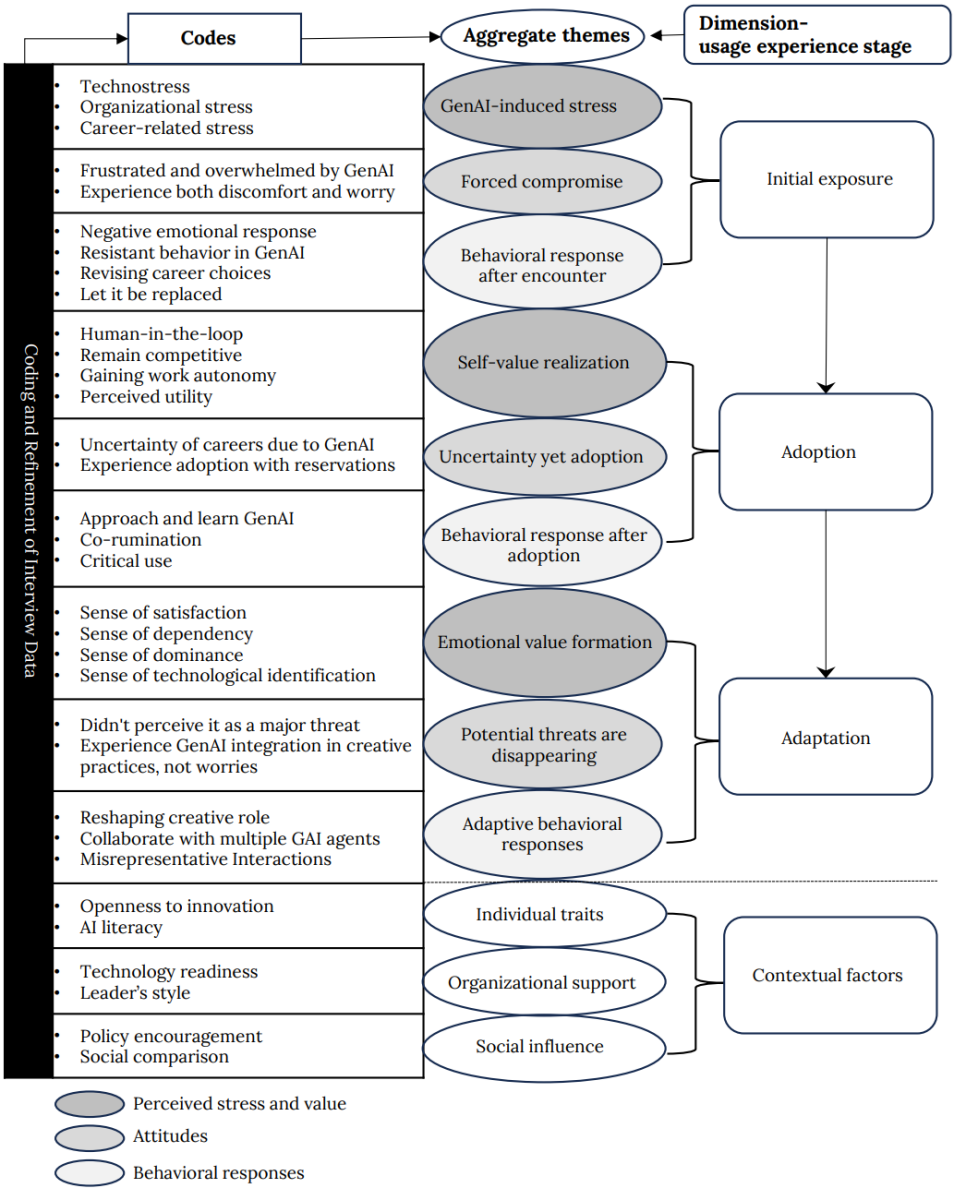


Figure 1. Themes and coding scheme

Findings

Findings are presented according to the twelve identified themes across participants' three usage experience stages.

Perceived stress and value in creative professionals

The coding analysis shows that creative professionals' engagement with GenAI involves initial exposure, adoption, and adaptive use. Figure 2 illustrates how participants' perceived stress and value evolve at different stages.

For participants in the ***initial exposure phase***, GenAI triggered significant stress, largely driven by occupational, technical, and organizational challenges. Occupational stress arises from GenAI lowering the creative threshold for professionals. P8 said, '*Almost anyone can use text-to-image AI to create art, the creative jobs might disappear*'. In addition, most participants reported that their creative occupation was classified as having high exposure to AI (Feltenet al., 2021) and was perceived as particularly vulnerable to AI replacement. While GenAI may be considered 'good

stress,' its technology advantages—such as strong fusion capabilities, anthropomorphic interactions, personalized services, rapid responsiveness, and high-quality content—could still threaten professionals, as inefficient AI might not be seen as a stressor. For example, P9 said, '*Why is AI advancing so rapidly? Because it truly learns from the data you give it, but people might not. It's quite frightening!*' Organizational stress stems from cost reduction and efficiency-driven, results-oriented culture. A marketing creator mentioned, '*For businesses, what once seemed like an impossible triangle (high quality, low cost, and short time) has become achievable with the advent of GenAI. We might not be able to reach it*' (P6).

During the ***adoption phase***, participants conveyed that perceptions of self-value realization, such as remaining competitive, gaining work autonomy, human-in-the-loop (HITL) and perceived utility, helped alleviate the initial three challenging stresses associated with GenAI. Compared to not adopting GenAI, embracing it can lead to greater work autonomy and to enhanced competitiveness (P1, P4, P11, P17). The need for human involvement in bridging GenAI's limitations strengthens the case for adopting GenAI, making it hard to move away from it under content creation stress. Similarly, the attraction of perceived utility seems to have weakened the stress on participants about content creative competition. As one participant noted, '*It's like an invisible assistant; I'm almost forgetting the career stress it caused*'(P10).

When participants enter the ***active adaptation stage***, adjusting and adapting to each other to GenAI will form their emotional value and subjectively enhance their sense of satisfaction, dependence, dominance, and technological identity, further influencing their initial perception of stress. For example, P15, a designer who has been using ChatGPT intermittently since its inception, shared: 'The *way GenAI and I can adapt to each other's needs is satisfying and removes any need for unnecessary concern.*' The sense of dominance stems from the guidance and mastery of the GenAI tool. For example, one participant stated, '*I can take control of GenAI and incorporate my ideas into the creation process to achieve my desired results*'(P7). Technological identity refers to an individual's identification with and sense of belonging to a particular technology. As one puts it,

> *AI, as a new quality productive force, has gradually reshaped the way I work in collaboration with it, and even if there is competitive stress, it is from other peers rather than AI* (P11).

Additionally, some participants acknowledged a growing dependence on GenAI during the adaptation phase, with a high likelihood of path dependency, which exacerbates the pressure to use the technology. For instance, P4 remarked,

> '*Before using ChatGPT, my work efficiency was 100%. After using it, it increased to 150% or even 200%. However, losing ChatGPT's assistance may drop my efficiency to 50% instead of returning to 100%.*'

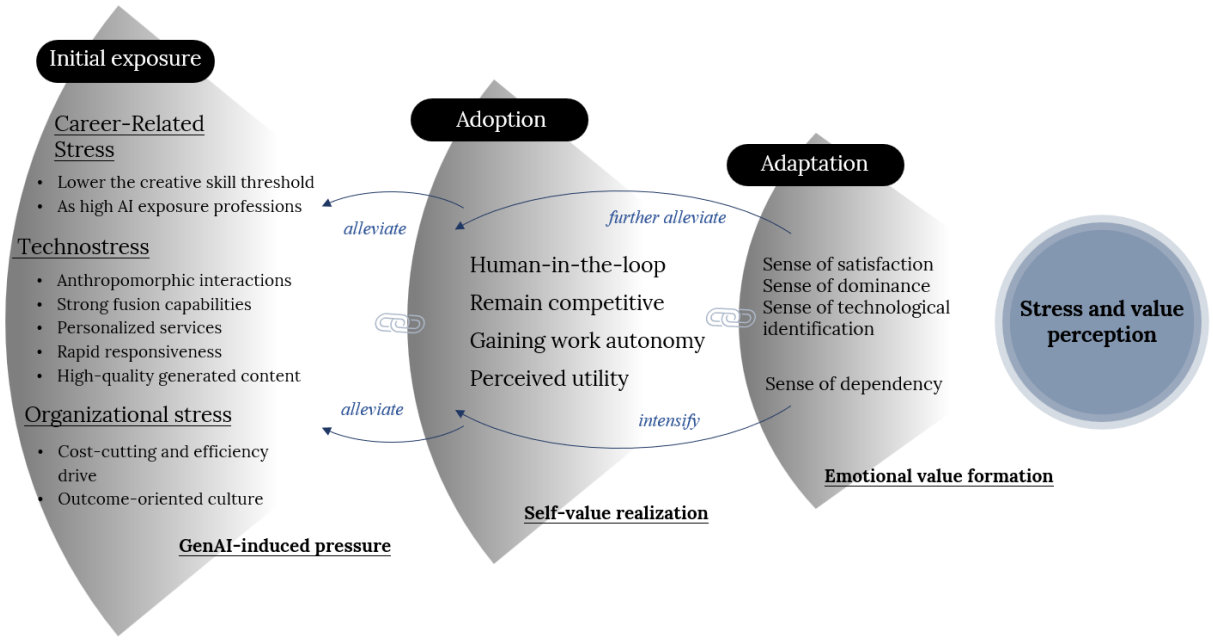


Figure 2. Stress and value perception in creative professionals

Attitudes and behavioral responses towards GenAI in creative professionals

Findings showed that creative professionals' attitudes toward GenAI in their work varied with their perceived value experiences at different stages of interaction. Participants' attitudes toward the use of GenAI may shift from initial worry to moderate concern, and eventually to less worry. These varying attitudes can trigger either positive or negative behavioral responses.

When participants were in the ***initial exposure stage***, they often had a strong negative attitude toward GenAI. At this stage, due to the predominant sense of threat and work stress forced creators in compromise, participants mainly exhibit negative maladaptive responses, including adverse emotional reactions, resistance behaviours, and negative career choices and adjustments. For example, one programmer (P18) expressed concern that the sense of crisis in his profession caused by GenAI led him to either silently or vocally resist it, but this resistance proved ineffective. Participants also conveyed that when they are particularly concerned that AI-generated content might undermine the value and significance of their creativity, they may exhibit resistance by withdrawing, avoiding, or opposing the use of GenAI in their workflow. Job burnout, leaving, and changing jobs were also behavioral responses mentioned repeatedly by participants in this phase. For example, P1 said, '*I used to be a programmer, and I quit because I was very aware of my career prospects. Sooner or later, my job would be replaced by AI.*'

In the ***adoption stage***, the stress of job replacement gradually transforms into a positive driving force for progress. Although creators adopt GenAI in their creative practices, the stress of competition can be activated at any time, causing them to fluctuate between an embracing attitude and concerns about their future careers. At this stage, some respondents take positive and constructive responses to navigate their challenges. They either adopt GenAI to assist with self-improvement and actively follow and recommend GenAI, or they begin to alleviate professional stress by continuously evaluating and critically reflecting on the use of GenAI. For example, the generation of mis/disinformation was identified by several respondents as one of the reasons they view GenAI with less tension. '*I find that ChatGPT often uses strange but plausible logic to arrive at conclusions that only seem correct at first glance*'(P2). Some participants (P2, P5, P12) cope with stress by over-discussing and sharing the negative experiences and flaws of GenAI in online communities, a behavior known as co-rumination (Hu et al., 2023).

The adoption of GenAI will further drive its exploratory and adaptive use (wael AL-khatib, 2023). At the ***adaptation stage***, some participants exhibit adaptive behavioral responses. Reshaping creative roles is notable, such as data analysts need to transition into '*AI model trainers and supervisors*'(P12), program developers perceive themselves no longer as mere '*code workers*' but as '*managers, controllers, and learners*' of the technology (P1), Photojournalism will undergo a role transformation from '*photographers to reporters to verifiers*'(P5). Another response is to collaborate with multiple GAI agents. A game music producer shared how they use AI tools in their workflow,

> *for lyrics, I use ChatGPT and NetEase Tianyin; for composing, Suno generates accompaniments and melodies; for vocals, ACE helps with voice replacement. For a complex music video, I use Kaiber, and for animations, I turn to Runway and Pika* (P9).

Finally, to protect themselves from job displacement as GenAI learns human behavior patterns, some creators deliberately mask their true needs during interactions with the technology. They use misrepresentative interactions to safeguard their job security and professional status. An advertising planner noted, '*I avoid sharing my expertise with ChatGPT because it might surpass my skills and render my role obsolete*'(P13).

Contextual factors in GenAI adaptation use

Participants' use of GenAI for their creative work depended on specific contexts individuals find themselves in, which encompassed three main factors: individual traits, organizational support, and social influence. These contexts may cause participants to linger in or move between different stages.

Findings on the individual traits of participants revealed that AI literacy and openness to innovation strongly influenced their stress, attitudes, and responses to GenAI. Those with higher AI literacy and a positive attitude towards innovation felt less competitive stress and were more excited than worried, demonstrating more positive adaptive coping behaviours. For example, P1, an illustrator, shared, '*I do meet many co-workers who are scared and resistant to it, but I embrace change and continuously adapt GenAI in my work*.'

If participants' individual traits enable them to proactively learn, adopt, and grow with GenAI naturally, then the organization and leadership will need them to do so. For example, P19 recalls her experience with GenAI: '*The company first encouraged independent use, then trained all staff on Midjourney and prompt skills, which I was initially reluctant to learn on my own*.' P14, in a leadership role, stated, '*I have mandated my employees to use AI for creation...*'

Social influence mentioned by participants, such as widespread support and promotion of GenAI in government policy and the media, also affects their engagement with GenAI. One mentioned, '*Government departments and big companies are already using AI; it's only a matter of time before it's everywhere. Embracing GenAI is a smart move*'(P16). Similarly, P17 observed, '*The new AI rules show that everyone's still backing the growth and use of GenAI*.'

Discussion and future research

This ongoing research uses exploratory qualitative methods to preliminarily explore creative professionals' perception and use of GenAI in content creation competitions with a value-attitude-behavior theory. Findings underscore that the transition to GenAI usage involves stages of initial exposure, adoption, and adaptation, highlighting the dynamic nature and self-adjustment of creative professionals. This contrasts with the static, cross-sectional findings on GenAI's impact and adoption. Our findings also address the call by Kshetri et al. (2023) to explore the career impacts of GenAI on creators from psychological, cognitive, and emotional micro-levels. Moreover, the different value perceptions, attitudes, and behavioral changes of creative professionals across the three stages suggest that the exposure effect (Wullenkord et al., 2016) is also present in the human-AI interaction between creative professionals and GenAI. Specifically, perceived stress can

pose a challenge to the effective use of GenAI, while in-depth experience and usage serve as a strategy to reduce the perception of GenAI as a source of stress or threat.

The practical implications of our findings include informing GenAI tool developers about targeting interventions for creative professionals at various stages of the design process. For instance, for creative professionals who are initially exposed to GenAI—those who view it with fear and stress—an ideal effective GenAI product should help them understand its underlying logic from the outset. It should build trust and promote informed engagement, rather than remaining a mysterious black box. As creative professionals become more familiar with and reliant on the tools during the adaptation phase, the GenAI system should help them maintain control over core creative tasks. It should ensure that GenAI not only supports the creative process but also provides creative professionals with greater autonomy in decision-making. Managers should guide employees in understanding GenAI's impact on their careers, enhance algorithmic and AI literacy training, and implement tailored management strategies to help creators adapt to GenAI integration, improving their sense of security and career belonging.

As part of an ongoing study, this paper has two main limitations: the sample does not include a diverse range of occupational levels, missing perspectives such as those from managers, and the findings are based on static self-report data, which could be biased; future research should use observational methods and multi-source data to provide a more objective analysis.

Acknowledgment

This research is supported by the Major Project of National Social Science Fund of China (No. 24&ZD180).

About the authors

Mingxia Jia is a Ph.D. candidate of Economics and Management at Nanjing University of Science and Technology. Her research interests include information behavior and human-computer interaction. She can be contacted at jiamx0301@njust.edu.cn.

Yuxiang Chris Zhao is a Professor of Information Management at Nanjing University in China. His research centres on human-computer interaction, health informatics, and social media. He can be contacted at yxzhao@vip.163.com.

Xiaoyu Zhang is a Ph.D. student at the University of Auckland, New Zealand. She focuses on human-AI collaboration and information behavior. She can be contacted at xyzhangis@outlook.com

References

Alon, L., & Krtalić, M. (2024, April). 'Words Are not just Words; They Carry Experiences Within Them': Navigating Personal Information Management in Multilingual Contexts. In *International Conference on Information* (pp. 333-342). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57850-2_25

Bankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D., & Woo, S. E. (2024). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, 45(2), 159-182. https://doi.org/10.1002/job.2735

Bannon, L. (2011). Reimagining HCI: toward a more human-centred perspective. *Interactions*, 18(4), 50–57. https://doi.org/10.1145/1978822.1978833

Boucher, J. D., Smith, G., & Telliel, Y. D. (2024, May). Is Resistance Futile?: Early Career Game Developers, Generative AI, and Ethical Skepticism. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-13). https://doi.org/10.1145/3613904.364188

Chang, P. C., Zhang, W., Cai, Q., & Guo, H. (2024). Does AI-Driven technostress promote or hinder employees' artificial intelligence adoption intention? A moderated mediation model of affective reactions and technical self-efficacy. *Psychology Research and Behavior Management*, 413-427. https://doi.org/10.2147/PRBM.S441444

Crabtree, B. F., & Miller, W. L. (1999). Clinical research: a multimethod typology and qualitative roadmap. *Doing qualitative research*, 2, 3-30.

De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, 7(11), 1845-1854. https://doi.org/10.1038/s41562-023-01734-2

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). Opinion Paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy. *International Journal of Information Management*, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12), 2195-2217. https://doi.org/10.1002/smj.3286

Granulo, A., Fuchs, C., & Puntoni, S. (2019). Psychological reactions to human versus robotic job replacement. Nature human behaviour, 3(10), 1062-1069. https://doi.org/10.1038/s41562-019-0670-y

Gupta, R., Nair, K., Mishra, M., Ibrahim, B., & Bhardwaj, S. (2024). Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. *International Journal of Information Management Data Insights*, 4(1), 100232. https://doi.org/10.1016/j.jjimei.2024.100232

Hu, B., Mao, Y., & Kim, K. J. (2023). How social anxiety leads to problematic use of conversational AI: the roles of loneliness, rumination, and mind perception. *Computers in Human Behavior*, 145, 107760. https://doi.org/10.1016/j.chb.2023.107760

Inie, N., Falk, J., & Tanimoto, S. (2023, April). Designing participatory ai: Creative professionals' worries and expectations about generative ai. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-8). https://doi.org/10.1145/3544549.3585657

Kim, M. J., Hall, C. M., Kwon, O., & Sohn, K. (2024). Space tourism: Value-attitude-behavior theory, artificial intelligence, and sustainability. *Journal of Retailing and Consumer Services*, 77, 103654. https://doi.org/10.1016/j.jretconser.2023.103654

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in human behavior*, 114, 106553. https://doi.org/10.1016/j.chb.2020.106553

Kshetri, N., Dwivedi, Y. K., Davenport, T. H., & Panteli, N. (2023). Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda. *International Journal of Information Management*, 102716. https://doi.org/10.1016/j.ijinfomgt.2023.102716

Kumar, A., Krishnamoorthy, B., & Bhattacharyya, S. S. (2024). Machine learning and artificial intelligence-induced technostress in organizations: a study on automation-augmentation paradox with socio-technical systems as coping mechanisms. *International Journal of Organizational Analysis*, *32*(4), 681-701. https://doi.org/10.1108/IJOA-01-2023-3581

Li, J., & Huang, J. S. (2020). Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. *Technology in Society*, 63, 101410. https://doi.org/10.1016/j.techsoc.2020.101410

Liang, X., Guo, G., Shu, L., Gong, Q., & Luo, P. (2022). Investigating the double-edged sword effect of AI awareness on employee's service innovative behavior. *Tourism Management*, 92, 104564. https://doi.org/10.1016/j.tourman.2022.104564

Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. International Journal of Qualitative Methods, 22, 16094069231205789. https://doi.org/10.1177/16094069231205789

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192. https://www.science.org/doi/10.1126/science.adh2586

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & mass communication educator*, 78(1), 84-93. https://doi.org/10.1177/10776958221149577

Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, 7(11), 1836-1838. https://doi.org/10.1038/s41562-023-01751-1

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1-11. https://doi.org/10.1038/s41562-024-01882-z

Vorobeva, D., El Fassi, Y., Costa Pinto, D., Hildebrand, D., Herter, M. M., & Mattila, A. S. (2022). Thinking skills don't protect service workers from replacement by artificial intelligence. *Journal of Service Research*, *25*(4), 601-613. https://doi.org/10.1177/10946705221104312

wael AL-khatib, A. (2023). Drivers of generative artificial intelligence to fostering exploitative and exploratory innovation: A TOE framework. *Technology in Society*, 75, 102403. https://doi.org/10.1016/j.techsoc.2023.102403

Woodruff, A., Shelby, R., Kelley, P. G., Rousso-Schindler, S., Smith-Loud, J., & Wilcox, L. (2024, May). How knowledge workers think generative ai will (not) transform their industries.

In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-26). https://doi.org/10.1145/3613904.364270

Wullenkord, R., Fraune, M. R., Eyssel, F., & Šabanović, S. (2016, August). Getting in touch: How imagined, actual, and physical contact affect evaluations of robots. In 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 980-985). IEEE. https://doi.org/10.1109/ROMAN.2016.7745228.

Yao, F., Li, C., Nekipelov, D., Wang, H., & Xu, H. (2024). Human vs. Generative AI in Content Creation Competition: Symbiosis or Conflict?. *arxiv preprint arxiv*:2402.15467. https://doi.org/10.48550/arXiv.2402.15467

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Taming TikTok: how BIPOC individuals perceive and interact with algorithmically generated content

Jiarun Dai, Naila Hajiyeva, Sehba Wan, and Kayla Booth

DOI: https://doi.org/10.47989/ir30iConf47089

Abstract

Introduction. TikTok's recommendation algorithm plays a crucial role in shaping user experiences, raising concerns about algorithmic bias, content suppression, and misinformation, particularly for BIPOC users. This study explores how BIPOC individuals perceive and interact with TikTok's algorithm, focusing on content visibility, algorithmic manipulation, and experiences with problematic content.

Method. This pilot study utilized semi-structured interviews with 10 BIPOC TikTok users, recruited via social media. Participants discussed their experiences with the platform's recommendation system, and responses were transcribed and thematically analyzed to identify patterns in content exposure, engagement strategies, and perceptions of algorithmic bias.

Analysis. A qualitative thematic coding approach was applied to categorize responses based on key themes, including content representation, misinformation, and strategies for mitigating algorithmic harm.

Results. Findings reveal that BIPOC users frequently encounter discriminatory content, misinformation, and algorithmic suppression. Many employ strategies such as selective engagement, the "Not Interested" feature, and algorithm resets to curate their feed. However, skepticism remains regarding TikTok's data practices and fairness in content moderation.

Conclusions. BIPOC users actively shape their digital experiences to counteract algorithmic bias, highlighting systemic inequities in AI-driven platforms. Future research should expand participant diversity and investigate long-term algorithmic trends to promote inclusive digital spaces.

Introduction

A recommendation algorithm is a system, classified as artificial intelligence (AI), that has the '*ability to predict whether a particular user would prefer an item or not based on the user's profile*' (Isinkaye, pg. 262). While these algorithms are embedded across a myriad of services, and platforms, there has been increased attention from scholars, popular press, and the US government on the recommendation algorithm that drives the TikTok application. At over one billion users worldwide, TikTok has grown faster than any application before it (Ungless, 2024). TikTok's fame for gaining such a large amount of users in a short amount of time, leads to questions regarding their practices with data, surveillance, and ethics, indicating the importance of understanding how this technology impacts societal norms, systems, and culture (Montag, 2021). According to a Pew Research Center survey, 62% of U.S. adults under 30 use TikTok, with the platform becoming an increasingly important source of news and political information for this demographic (Eddy, 2024). Dissecting TikTok's popularity and high user amount, is vital to understanding the current systems of recommendations.

Although algorithms shape the usage of TikTok, users have learned how to gamify the app to manipulate their For-You-Page, which is an endless stream of videos personalized for each user (Bhandari, 2022). Algospeak, known for abbreviating, misspelling, or substituting specific words to avoid content moderation, is a practice that not only general users engage in, but marginalized users are familiar with as they deal with content review, and appeal processes where their content is being deemed as '*inappropriate*' when it is not (Steen, 2023). While there are a variety of definitions for who is considered '*marginalized*', the National Library of Medicine defines marginalized people '*as individuals who are excluded from mainstream social, economic, media or cultural life,*' (National Library Medicine). For autistic individuals, a marginalized community that is typically misrepresented, digital media has allowed them to provide counter-narratives to dehumanizing perceptions surrounding autism (Alper, 2023). However, on TikTok, it can be difficult to find informational and accurate videos as 41% of videos were classified as inaccurate when searching the '*#Autism*' hashtag (Aragon-Guevara, 2023). Content suppression undermines the inclusivity that marginalized users are fighting for in digital spaces.

Ruha Benjamin refers to this oppression as the '*New Jim Code*'. This code emphasizes how new technologies are deemed as '*progressive*' but still push forth the same social inequalities (Benjamin, 2020). While marginalized users are known to benefit uniquely from this platform, without uprooting these technical systems and questioning who benefits from them, marginalized folks will continue to experience discrimination and intolerance (Antoine, 2022). BIPOC communities are experiencing a lack of understanding of their user experiences of TikTok, where biases and stereotypes should be acknowledged. In Algorithms of Oppression, Noble talks of hiring software creating inequities within the job hiring process. Bursell et al, conducted an experiment testing AI hiring software, discovering a bias towards white middle-aged women. This showcases that recommendation algorithms are playing a role in social exclusion, inequality, and discrimination by algorithms (Bursell, 2024). By easily accepting these algorithms, individuals are allowing bias to enter digital spaces and cause harm that may go unnoticed (Timmons, 2021).

In this paper, we examine the intersection of BIPOC identity and TikTok's recommendation algorithm to better understand its impact on marginalized users' experiences and potential algorithmic harms. This short paper is dedicated to a preliminary and exploratory study that begins to fill the gaps identified in the literature review. We're piloting semi-structured interviews focusing on identity, experiences with TikTok, and strategies for handling harmful content. Through thematic analysis of interview transcripts, our findings reveal insights into how these BIPOC users navigate TikTok's algorithm-driven environment, including methods for managing exposure to harmful content and adapting to problematic algorithmic behaviours. Our research provides valuable perspectives on the challenges faced by marginalized users in algorithm-centric

platforms, offering insights into creating equitable and inclusive digital spaces in an increasingly AI-mediated social landscape.

Literature review

Our literature review builds upon a systematic review conducted in our previous poster paper, which examined research trends at the intersection of BIPOC identity and recommendation algorithms in HCI. We extend this work by exploring algorithmic literacy, marginalized users' experiences, and BIPOC user experiences specifically on TikTok.

Algorithmic literacy

Algorithmic literacy is the term used to describe the capacity and opportunity to be aware of both the presence and impact of algorithmically driven systems (Oeldorf-Hirsch, 2023). Personalization can be defined as mechanisms to customize information and content to users (Perugini). Within personalization, building community is a key concept of TikTok's aims to create permeable communities - communities one can weave in and out of - that allows users to discover not only their identities but others (Milton). This has its disadvantages as unregulated discoverability can showcase harmful content to users. However, personalization showcases an advantage to those who understand how to use the application, as they grasp the effects of curating their own experience, rather than just letting the algorithm adapt. Aside from algorithmic literacy, the curation element of TikTok is a reflection of navigation as the user has agency in what is presented to them - to an extent. With recommendation algorithms being based off of a user's interaction such as search, should the For-You-Page reflect those searches is a question raised by Milton et al about the '*feed*' of TikTok. There are concerns about whether or not searches are a reliable source of information about a user.

Marginalized users and experiences with recommendation algorithms

For marginalized users, content review and moderation is a large element of what their algorithmic experiences are. Shadow banning is a practice of removing videos without a user's knowledge, which disproportionately impacts marginalized users (DelMonaco). Rong et al discusses how this tactic impacts blind users often suppressing their content to a more mainstream audience, and how they find themselves by trying to immerse themselves as much into staying in their desired communities (Smith, 2023).

To combat content suppression and mitigation there are a number of techniques that marginalized users partake in. A common method is avoiding hashtags as that spreads the material far out of the creators' typical interactions and communities (DelMonaco). Another method is sticking to trending topics as there is a large saturation of material, so the content does not always spread as much.

BIPOC users and experiences with recommendation algorithms

BIPOC users are becoming more of a voice in bringing awareness to the harms of TikTok, both technically and socially. Williams conducted a study where Black creators shared content mitigation approaches to their interaction within the application. There is limited research on the nature of content BIPOC users encounter and how they interact with potentially problematic content. Some studies have examined specific content types such as when Hung et al (2023) conducted content analysis on videos tagged '*#racialtrauma*'. The team found a frequent theme of '*questioning and challenging oppression, denial and privilege*', where creators drew comparisons of their own experience against other creators (Hung, 2021). This theme reflects on the experience of BIPOC individuals on TikTok, where they question and discuss their own experiences surrounding racism. Existing research has concentrated on content creators, particularly how BIPOC creators deal with content suppression and algorithmic bias. Although the data from this study was not discussed in the context of mental health or feelings surrounding content by BIPOC

creators, it is important to note how BIPOC creators focus on education and advocacy through their content, to fight algorithmic bias.

Gaps in the literature

Throughout all the literature discovered, there has yet to be a focus of BIPOC user experiences. There's limited research on the nature of content BIPOC users encounter, especially on how they perceive and interact with potentially problematic content. While some studies have examined specific content types (e.g., Hung et al.'s study on '*#racialtrauma*' videos), there's a gap in understanding the broader content landscape for BIPOC users. A majority of the existing research has concentrated on content creators' experiences, particularly how BIPOC creators deal with content suppression. There's less information about the experiences of BIPOC individuals who primarily consume content on TikTok. Improving the TikTok experience is vital to creating an inclusive platform, as the neglect to amplify voices that should be seen reflects a failure to listen to those who deserve to be heard (Antoine, 2022).

Methodology

This pilot study utilizes semi-structured interviews to explore the nuanced experiences of BIPOC TikTok users, allowing us to capture the complexity of their perspectives. Participants were provided with detailed information about the study, providing their informed consent before participating. Measures were taken to ensure the confidentiality of participants' data, including anonymizing transcripts and securely storing recordings.

Research questions:

1. What is the nature of content BIPOC users encounter on their For You Page (FYP)?
2. How do BIPOC users perceive and interact with problematic content on TikTok?
3. What are BIPOC users' perceptions and understanding of TikTok's recommendation algorithm?
4. In what ways do BIPOC users attempt to influence the TikTok algorithm to curate their content experience?

Recruitment

The participants for the pilot study were selected based on specific inclusion criteria to ensure a diverse and relevant sample. We conducted recruitment by publishing a flyer, in the i3 Instagram and Slack, with the criteria titled '*Research Opportunity: We Want to Hear About your TikTok Experience*'. In the flyer we briefly explained the purpose of our study, inviting readers to join us in centring the voices of BIPOC users of TikTok. Eligibility criteria included:

1. BIPOC individual
2. Age between 18 and 40
3. A TikTok user
4. English speaker

The flier included a link to a Google form for potential participants to provide their contact information and confirm their eligibility. Emails were scheduled by giving each participant a few options for time slots. However, as we expand the number of participants, we will be using Calendly to make it easier for them to schedule time slots that work for our schedule as researchers. This flyer was available on i3's Instagram and Slack channels.

Data collection

Interviews were scheduled conveniently for participants and interviewers and conducted on Zoom. Upon conducting 10 semi-structured interviews, we transcribed the interviewers' answers and put them in a Google Sheet File, split by question. Each interview lasted between 20 and 40

minutes. All the interviews were audio recorded with participants' consent to ensure accurate transcription and analysis.

Data analysis

We transcribed the interviewers' answers into an organized Google Sheet file, split by question. We fixed the transcription errors and only kept the complete answers from the participants, noting our follow-up questions. Then, each of us conducted qualitative coding, starting by writing out the initial themes for answers from all 10 participants. We included quotes that stood out to us in their answers, including a list of recurring themes. Then, we summarized those into focused codes - themes we've seen across multiple answers to the same question. Then, those focused codes were combined between answers to summarize findings for each question from the interview pool. For every response, we noticed similarities and themes across the questions to combine the topics and restructure the interviews. As is shown in Table 1, the age range of our participants varied between 19 and 27 instead of including adults in their 30s. This affected the scope of our study and only made the results applicable to the age group interviewed.

Participant #	Gender	Race	Participant Age
1	Female	Latino	24
2	Female	Latino	21
3	Female	Asian	21
4	Female	Asian	21
5	Female	Latino	27
6	Male	Black	22
7	Male	Latino	20
8	Female	Latino	20
9	Male	Latino	21
10	Female	Black	19

Table 1. Demographic information of interview participants

Results

Nature of content BIPOC users encounter on their for You Page (FYP)(RQ1)

We examined the types of content BIPOC users encounter on their TikTok FYP (Table 2). The most frequently encountered content type was BIPOC Representation, followed by Humor and Entertainment, and Cultural Relevance. While this diversity suggests a broad range of content, the relative scarcity of educational or critical discourse content may indicate an algorithmic preference for entertainment over information.

Types of Content	# of Participants
BIPOC Representation	9
Humor and Entertainment	8
Cultural Relevance	7
Educational Content	4
Content From Friends/Acquaintances	3

Table 2. Types of content encountered on FYP

BIPOC users' perceptions and interactions with problematic content (RQ2)

Participants identified problematic content, including misinformation, discriminatory material, and unintentional stereotypes. Their responses included disengagement, algorithm manipulation, and information-seeking. However, the effectiveness of these responses remains uncertain.

Misinformation

Half of the participants (5 out of 10) reported encountering misinformation, particularly regarding health. Participant 4 noted:

> *I have seen a lot of misinformation, especially health on TikTok. I've seen a lot of people saying 'Oh, do this, and you'll lose 50 pounds', or 'take this pill, and this will happen to you'* (Participant 4).

This underscores the prevalence of misleading health advice, which is particularly concerning for BIPOC users facing healthcare disparities.

Discriminatory content

A majority (6 out of 10) encountered discriminatory or stereotypical content. Participant 10 shared frustration: '*There are pages that are full of it. It's just like, that's a lot. You need to take a breather*' (Participant 10). This suggests that such content is pervasive enough to significantly impact user experience.

Unintentionally harmful content

Four participants noted content that, while not overtly discriminatory, perpetuated stereotypes. Participant 2 observed: '*I definitely don't think it's intentional. I think there's just a lot of stereotypes*' (Participant 2). This highlights the challenge of addressing unintentional bias in content creation.

Responses to problematic content

Participants employed various strategies:

1. Disengagement: the most common response (8 out of 10 participants) was to disengage from problematic content. Participant 6 explained:

> *I try to understand maybe they're just doing it for views for content. They're trying to get a reaction so I just kind of just be like, 'well it's not really my cup of tea, so I'll just skip to the next video* (Participant 6).

2. Reporting: a minority of participants (3 out of 10) mentioned actively reporting content they found problematic.

3. Algorithm manipulation: many participants (7 out of 10) used features like "Not Interested" to influence their algorithm. Participant 10 shared: '*I love to press 'not interested'. And also, just swiping away. I don't like, don't say, don't comment, don't favourite*' (Participant 10).

4. Information seeking: some participants (2 out of 10) engaged further by reading comments or seeking more information. Participant 3 noted:

> *I'll just interact with them, anyway. I definitely have commented on videos like that before just giving my personal opinion. But what I'll mostly do is look for more information.* (Participant 10).

These findings suggest that BIPOC users actively curate their experience and employ various strategies to mitigate exposure to problematic material.

BIPOC users' perceptions and understanding of TikTok's recommendation algorithm (RQ3)

Participants demonstrated an understanding of TikTok's algorithm, emphasizing engagement-based recommendations, data collection, and cross-platform influence. While most accepted this as inherent to social media, concerns over privacy and content manipulation were evident.

Engagement-based understanding

Nearly all participants (9 out of 10) believed that user engagement drives recommendations. Participant 9 explained:

> *I imagine a system behind the scenes that looks into how long you look at a certain type of video, what videos you like, what videos you bookmark and gives you recommended videos based on those stats* (Participant 9).

This indicates a sophisticated grasp of how modern recommendation systems operate.

Data collection awareness

Many (7 out of 10) were aware of TikTok's extensive data collection. Participant 5 remarked:

> *I don't convince myself that there's any resemblance of data privacy or security anymore. I know that whatever powers that be that control the things in my phone are stealing my data, and there's very little I can do about it* (Participant 5).

This reflects a resigned attitude towards data privacy.

Cross-platform influence

Half (5 out of 10) suspected TikTok recommendations were influenced by activities on other platforms. Participant 4 noted: '*I do have some suspicions that it also has access to other information. Maybe, like what I'm searching for on other platforms*' (Participant 4).

Location-based content

Some (4 out of 10) observed location-based content recommendations. Participant 7 explained: '*It definitely was based on my location at first, so I feel like just to get a basis on people*' (Participant 7).

Content manipulation beliefs

A minority (3 out of 10) believed TikTok manipulates content visibility. Participant 10 shared:

> *I think that specifically it shows, the most on the for you pages also when, like your analytics in terms of content, creators how their videos say things like shadow banned or the video within a for you page is pushed by what you as a like user and interact with the most* (Participant 10).

BIPOC Users' attempts to influence the TikTok algorithm (RQ4)

BIPOC users actively curate their content by engaging with preferred material, avoiding unwanted content, and periodically resetting the algorithm. However, these strategies suggest a reactive rather than an empowered approach.

Active engagement

Most participants (8 out of 10) engaged with desired content to shape their feed. Participant 2 noted:

> *I usually just, once I get to a video, I like or bookmark it. I have tried it to see if my theory is sometimes right. So, I just either keep liking videos, or commenting, or engaging with them for a long time* (Participant 10).

Searching for desired content

Many (6 out of 10) used search functions to refine their FYP. Participant 9 shared:

> *So, there's a show. I watch and I wanna see more content about it since I get it all on twitter. So, I started looking it up on Instagram or on TikTok. And I, my feed did get populated with more of those videos and memes around the show* (Participant 9).

Disengagement from unwanted content

Almost all participants (9 out of 10) actively disengaged from undesired content, often using the '*Not Interested*' feature.

Periodic algorithm reset

A few (2 out of 10) intentionally reset their algorithm. Participant 3 described: '*so, I actually tried to refresh my algorithm one day because I was just curious. And I just kept scrolling one video after the other*' (Participant 3).

Cross-platform manipulation

Some (3 out of 10) attempted to influence their TikTok feed through other platforms. Participant 5 noted: '*If I just Google things on my phone, my ads on TikTok will eventually change*' (Participant 5). These findings demonstrate that BIPOC users are not passive consumers but active participants in shaping their digital experience, using various strategies to curate their FYP and tailor their TikTok experience to their preferences.

Discussion

Content curation and representation

The high prevalence of entertaining (8/10 participants) and culturally relevant content (7/10 participants) on BIPOC users' FYPs demonstrates TikTok's effectiveness in curating personalized content, aligning with research on social media algorithms' ability to create tailored content ecosystems (Eg et al., 2023). The significant presence of BIPOC-centered content (9/10 participants) suggests success in promoting diversity and addressing historical underrepresentation in mainstream media (Titanji et al., 2022). However, the emphasis on entertainment over education (4/10 participants) raises questions about meaningful knowledge dissemination.

Navigating problematic content

All participants reported encountering problematic content, with a preference for disengagement (8/10) and overactive reporting (3/10). This behavior aligns with research on digital resilience strategies among marginalized communities (Du et al., 2020) and highlights ongoing content moderation challenges. However, this may not be effective in reducing exposure over time. Future studies could examine whether TikTok's algorithm adjusts based on disengagement, or if problematic content persists despite user efforts.

User agency and algorithm awareness

BIPOC users demonstrate high algorithm awareness and actively attempt to influence their content recommendations, challenging notions of algorithmic determinism and emphasizing user agency in shaping digital experiences. The findings highlight an ongoing tension between personalization and control. While users employ strategies to influence content recommendations, they also express scepticism about TikTok's neutrality. Theories of algorithmic bias (Noble, 2018) could provide a useful lens to further analyze these concerns.

Limitations and future work

Expanding participant diversity

This study included 10 participants, offering in-depth but limited perspectives. Future research should aim for a more diverse sample of BIPOC users, considering intersectional factors such as age, gender, socioeconomic background, and specific ethnic identities. This approach would provide a richer understanding of how different BIPOC subgroups interact with and perceive TikTok's algorithm, acknowledging the complexity of intersecting identities and experiences

Conclusion

User experiences with social media are vital as applications such as TikTok are an essential part of culture now. As we navigate '*Living in an AI-algorithmic World*,' understanding how BIPOC users interact with and are influenced by algorithm-driven platforms like TikTok becomes increasingly crucial. HCI scholars have been increasing scholarly material on the impact of this technology, but little is known about the user experiences of BIPOC individuals. In this paper, we present an interview study of 10 BIPOC TikTok users. We find that content mitigation is crucial to navigating online experiences and that users tend to stick to their communities while acting accordingly when dealing with permeable communities. We also discuss the findings with the literature on relevant topics and talk about how our findings can be used to consider overall TikTok user experiences. These insights contribute to our understanding of how diverse user groups engage with AI-driven content ecosystems, reflecting the broader implications of living in a world shaped by algorithmic processes and generative AI technologies. This research directly addresses the iConference 2024 theme of '*Living in an AI-algorithmic world*' by examining how BIPOC individuals navigate TikTok's AI-driven recommendation system. By centring BIPOC perspectives, this study contributes to ongoing efforts to create more equitable and inclusive AI-driven platforms, ensuring diverse user needs are prioritized as we continue to navigate our AI-algorithmic reality.

Acknowledgements

We would like to sincerely recognize and thank the iSchool Inclusion Institute (i3), generously funded by The Andrew W. Mellon Foundation, for their valuable feedback and support on this work. Their guidance has been instrumental in shaping our research and empowering us to move forward. We are also deeply grateful to our research advisor, Dr. Kayla Booth, for her unwavering support and mentorship throughout this process. Her insights and encouragement have been invaluable. Additionally, we extend our heartfelt appreciation to all i3 scholars. Without their collaboration and dedication, this publication would not have been possible. This achievement holds great significance for us, and we sincerely thank everyone who has been a part of this journey.

About the authors

Jiarun Dai recently graduated with a bachelor's degree in Economics and Information Science from the University of Wisconsin–Madison. She will be pursuing a Ph.D. in Health Informatics. Her research interests include AI in healthcare, patient portals, and big data applications in medical decision-making. She can be contacted at jrddai1216@gmail.com

Sehba Wani is a recent graduate of the Information College at the University of Maryland, College Park. Her research interests involve sociotechnical systems, privacy, and information visualization. You can contact her at swani@terpmail.umd.edu

Naila Hajiyeva is a recent graduate from UT Austin with a Master of Science in Information Studies and a Bachelor of Science in Mathematics. Her research interests are based in urban science and social media algorithms. Naila can be contacted at naila.hajiyeva@utexas.edu

Kayla Booth is an Assistant Professor in the School of Information at The University of Texas at Austin. She is the Principal Investigator and Director of the iSchool Inclusion Institute (i3), a national undergraduate research and leadership development program. She received her Ph.D. from Penn State University, and her research focuses on the co-design and evaluation of community-led health and education interventions to support marginalized youth populations. She can be contacted at kayla.booth@ischool.utexas.edu

References

Alper, M., Rauchberg, J. S., Simpson, E., Guberman, J., & Feinberg, S. (2023). TikTok as algorithmically mediated biographical illumination: Autism, self-discovery, and platformed diagnosis on #autisktok. *New Media & Society*. https://doi.org/10.1177/14614448231193091

Antoine, A. (2023). 'We're Not Just We Not Just Well Suited, Well Suited, we are Ethically Obligated!' e Ethically Obligated!": Reimagining Technical and Professional Communication Pedagogy as Advocacy Against Discriminatory Digital Technologies. *Electronic Theses and Dissertations*, 2020-2023, 1174.

Benjamin, R. (2020). *Race after technology: Abolitionist Tools for the new jim code*. Polity.

Bhandari, A., & Bimo, S. (2022). Why's everyone on TikTok now? the algorithmized self and the future of self-making on social media. *Social Media + Society*, 8(1), 205630512210862. https://doi.org/10.1177/20563051221086241

Bursell, M., & Roumbanis, L. (2024). After the algorithms: A study of meta-algorithmic judgments and diversity in the hiring process at a large multisite company. *Big Data & Society*, 11(1). https://doi.org/10.1177/20539517231221758

Bustamante, C. M. V. (2022). Technology Primer: Social Media Recommendation Algorithms. (A. Higuchi & A. Jayanti, Eds.). *Technology and Public Purpose Project, Harvard Kennedy School*.

Du, J. T., Xie, I., & Waycott, J. (2020). Marginalized communities, emerging technologies, and social innovation in the digital age: Introduction to the special issue. *Information Processing & Management*, 57(3), 102235. https://doi.org/10.1016/j.ipm.2020.102235

Eddy, K. (2024, April 3). *6 facts about Americans and TikTok*. Pew Research Center. https://www.pewresearch.org/short-reads/2024/04/03/6-facts-about-americans-and-tiktok/

Eg, R., Demirkol Tønnesen, Ö., & Tennfjord, M. K. (2023). A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. *Computers in Human Behavior Reports*, 9, 100253. https://doi.org/10.1016/j.chbr.2022.100253

Harris, C., Johnson, A. G., Palmer, S., Yang, D., & Bruckman, A. (2023). 'honestly, I think TikTok has a vendetta against black creators': Understanding black content creator experiences on Tiktok. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 1–31. https://doi.org/10.1145/3610169

Hung, Y.-H., Miles, A., Trevino, Z., DAniello, C., Wood, H., Bishop, A., & Monshad, Z. (2023). BIPOC experiences of racial trauma on TikTok: A qualitative content analysis. *Contemporary Family Therapy*, 45(3), 298–308. https://doi.org/10.1007/s10591-023-09669-6

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3), 261–273. https://doi.org/10.1016/j.eij.2015.06.005

Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, 8(4). https://doi.org/10.14763/2019.4.1424

Lee, A. Y., Mieczkowski, H., Ellison, N. B., & Hancock, J. T. (2022). The algorithmic crystal: Conceptualizing the self through algorithmic personalization on TikTok. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–22. https://doi.org/10.1145/3555601

Lewis, J. A. (2024, March 13). *Tiktok and National Security*. CSIS. https://www.csis.org/analysis/tiktok-and-national-security

McClain, C. (2024a, August 20). *About half of TikTok users under 30 say they use it to keep up with politics, news*. Pew Research Center. https://www.pewresearch.org/short-reads/2024/08/20/about-half-of-tiktok-users-under-30-say-they-use-it-to-keep-up-with-politics-news/

Montag, C., Yang, H., & Elhai, J. D. (2021). On the psychology of TikTok use: A first glimpse from empirical findings. *Frontiers in Public Health*, 9. https://doi.org/10.3389/fpubh.2021.641673

Oeldorf-Hirsch, A., & Neubaum, G. (2023). What do we know about algorithmic literacy? the status quo and a research agenda for a growing field. *New Media & Society*. https://doi.org/10.1177/14614448231182662

Rong, E. Z., Zhou, M. M., Lu, Z., & Fan, M. (2022). 'it feels like being locked in a cage': Understanding blind or low vision streamers' perceptions of content curation algorithms. *Designing Interactive Systems Conference*. https://doi.org/10.1145/3532106.3533514

Sevelius, J. M., Gutierrez-Mock, L., Zamudio-Haas, S., McCree, B., Ngo, A., Jackson, A., Clynes, C., Venegas, L., Salinas, A., Herrera, C., Stein, E., Operario, D., & Gamarel, K. (2020). Research with marginalized communities: Challenges to continuity during the COVID-19 pandemic. *AIDS and Behavior*, 24(7), 2009–2012. https://doi.org/10.1007/s10461-020-02920-3

Siles, I., Valerio-Alfaro, L., & Meléndez-Moran, A. (2022). Learning to like TikTok . . . and not: Algorithm awareness as process. *New Media & Society*, 146144482211389. https://doi.org/10.1177/14614448221138973

Smith, H. (2023). Algorithms and the alphabet mafia: how TikTok influenced gender, sexuality, and the LGBTQ+ community during the COVID-19 pandemic. *University of Tennessee at Chattanooga, UTC Scholar*.

Steen, E., Yurechko, K., & Klug, D. (2023). You can (not) say what you want: Using Algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media + Society*, 9(3). https://doi.org/10.1177/20563051231194586

Timmons, K. (2021). Pre-Employment Personality Tests, Algorithmic Bias, and the Americans with Disabilities Act. *Penn State Law Review*, 125(2). Available at: https://elibrary.law.psu.edu/pslr/vol125/iss2/2

Titanji, B. K., Abdul-Mutakabbir, J. C., Christophers, B., Flores, L., Marcelin, J. R., & Swartz, T. H. (2022). Social Media: Flattening hierarchies for women and black, indigenous, people of colour (BIPOC) to enter the room where it happens. *Clinical Infectious Diseases*, 74(Supplement_3). https://doi.org/10.1093/cid/ciac047

Ungless, E. L., Markl, N., & Ross, B. (2024). Experiences of Censorship on TikTok Across Marginalised Identities. *Proceedings of the International AAAI Conference on Web and Social Media*. https://doi.org/10.48550/arXiv.2407.14164

Williams, G. (2023). Black Content Creators' Responses and Resistance Strategies on TikTok. *arXivLabs*. https://doi.org/10.48550/arXiv.2312.12727

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Examining generation Z's use of generative AI from an affordance-based approach

Chei Sian Lee, Li En Tan, and Dion Hoe-Lian Goh

DOI: https://doi.org/10.47989/ir30iConf47083

Abstract

Introduction. This paper uses the affordances framework to investigate how Generation Z (GenZ) students in higher education use generative AI (GenAI). There is an increasing need to gain a deeper understanding of GenZ's interaction with artificial intelligence tools to better support their integration into higher education and the workforce.

Method. Data was collated from semi-structured interviews with 34 GenZ students in higher education.

Analysis. Thematic analysis was conducted on the qualitative data collected from the semi-structured interviews.

Results. The findings suggest GenZ students have seamlessly integrated GenAI into diverse aspects of their lives. This study highlighted three main GenAI affordances that resonate with GenZ students: a) content searching and curation b) content generation and ideation, and c) content enhancement and refinement, revealing new opportunities for information access.

Conclusions. This study shed light on the perceived affordances of GenAI for GenZs, addressing a gap in the current literature on GenAI. The findings underscore the significant extent to which GenAI has been integrated into GenZ students' daily lives. Our study contributes to a better understanding of how GenAI's affordances facilitate and support GenZ students, providing invaluable insights that can inform future policies on developing literacy for AI use tailored to this group.

Introduction

Generation Z (GenZ), born between 1997 and 2012, now forms the bulk of undergraduate cohorts (Seemiller & Grace, 2017), and older GenZs have even begun to enter the workforce. As the first generation to grow up with a strong dependency on digital technology, information practices differ greatly from previous generations. (Lee et al., 2023b). Moreover, the COVID-19 pandemic coincided with their formative years, driving an accelerated adoption of technology in both educational and professional environments (Howard et al., 2023). Hence, there is an increasing need to gain a deeper understanding of GenZ's interaction with emerging technologies, such as artificial intelligence, to better support their integration into higher education and the workforce.

With the growing prevalence of generative artificial intelligence (GenAI) in education and the workplace, GenAI has become an integral component of GenZ's everyday technological resources. Specifically, GenAI affords a wide range of resources enhancing creativity, productivity, and decision-making across various contexts in their everyday life (Hirvonen et al., 2023). Prior research has shown that GenZs engage in higher levels of social activity on digital platforms and derive greater emotional satisfaction from digital media than older generations (Zhitomirsky-Geffet & Blau, 2017). Since GenZ is already accustomed to using digital media, GenZ students' dependency on resources afforded by GenAI is likely to be further amplified. (Taecharungroj, 2023). Developing a holistic understanding of how GenZ students in higher learning interact with GenAI is critical for optimizing education, preparing the future workforce, and ensuring AI tools are leveraged in ways that align with their needs.

In terms of affordance-based research, affordances are manifested in the form of distinctive technological attributes that shape the nature of communication (Lee, 2010; Sundar et al., 2015). Hence, we focus on features of GenAI deemed critical for everyday learning and interaction among GenZ students in higher learning. For instance, GenAI supports skills development (e.g. coding, writing, designing) that aligns with GenZ's tech-savvy nature. Further, GenZ students can use GenAI to brainstorm ideas, develop creative projects, and expand their critical thinking (Chan & Lee, 2023).

Within the information science domain, affordances are viewed as '*information opportunities*'—opportunities that allow individuals to engage in particular actions, thereby enabling access to information (Hirvonen et al., 2023; Lloyd, 2005). However, empirical research from such an information-oriented perspective on GenAI remains scarce. To fill this gap, this study aims to understand the affordances of GenAI that appeal to GenZs from an information-oriented perspective, focusing on how information is searched, accessed, and utilized. In the context of this study, we argue that GenAI affordances, as perceived by GenZ, create opportunities for efficient information access, shaping information practices that enable them to complete tasks more effectively.

Related work

Generation Z's use of digital technology

There are several factors influencing GenZ's information-seeking and learning processes on digital platforms. Firstly, GenZs are well-connected and go online regularly, due to their FOMO (fear of missing out) mindset (Nicholas, 2020). This mindset pushes them to constantly search for new information and to stay informed about trends, news, and social activities, whether related to personal interests, academic topics, or broader societal events (Lee et al., 2023b). Compared to previous generations, GenZ has had more access to the internet and digital media and, consequently, more access to information. Hence, GenZs are more likely to have a higher confidence level in navigating digital technology, and to prefer digital mediums for learning (Mosca et al., 2019). Furthermore, they have developed unique learning preferences as a result: they are more inclined to independent learning due to the individual nature of technology and prefer learning opportunities where they can immediately apply what they learn to real life (Seemiller & Grace, 2017; Lim & Lee, 2024).

Secondly, GenZs process information differently from previous generations. They have shorter attention spans (Giunta, 2017) and prioritize speed over accuracy (Nicholas, 2020). This could be attributed to the use of high-speed technology during formative years (Medina, 2008). As such, delivering information through concise content will likely be more effective at engaging GenZs (Tomaszewski, 2023).

Thirdly, many GenZs are currently navigating transitions to the next stage of education or the workforce (Lee et al., 2024). Individuals undergoing transitions engage in various practices to overcome challenges in the new information environment, such as calibrating their behavior by comparing with local practices and collecting important information (Hicks, 2022, Mulaudzi, 2023). For example, social media can provide the resources needed for calibration, by offering a glimpse into university social life (Dennen & Bagdy, 2024).

While research about GenZ's relationship with digital technology is abundant, there is a lack of literature specifically addressing their use of GenAI, largely because it is relatively new.

Affordances of generative AI

The affordances of an environment are what it offers the user, either for good or for ill (Gibson, 1986). Affordances are relational and exist in the interaction between an individual's subjective perception of utility and objective qualities of a technology (Schrock, 2015; Lee, 2010). Individuals perceive the functions and constraints of the technology, possess different levels of capability to use the technology, and interact differently with the technology depending on cultural norms and institutional regulation (Davis, 2020).

The theory of affordances highlights that technologies request, demand, encourage, discourage, refuse and allow certain actions and social habits (Davis, 2020; Lee et al., 2023a). The development of GenAI has led to a new generation of affordances for everyday information-seeking. For example, GenAI encourages users to seek information by generating it or modifying existing content and discourages users from comparing multiple sources (Hirvonen et al., 2023). GenAI also has numerous learning affordances. It allows anyone to generate text without deep understanding of the topic, allows readers to modify texts to different genres or styles, and provides writing feedback that could be used to consolidate learning (Warschauer et al., 2023). Examining interactions of people and GenAI through the lens of affordances can help us understand how the technology enables and constrains their information practices, and how GenAI will shape information practices towards a direction to complete an information-seeking task. In particular, this study focuses on understanding GenAI affordances that resonate with GenZ students and their potential implications.

Methodology

A total of 34 GenZ students entering university were recruited in Singapore to participate in a semi-structured interview. Of these, 47% of participants were male (n=16), and 53% (n=18) were female. Participants' ages ranged from 19 to 23. Data for this study were part of a larger study to examine Generation Z's information practices and the digital media they adopt in their transition to university. To incentivize participation, each participant was given S$25 SGD after completion.

Each interview lasted approximately 45 to 60 minutes and was audio recorded and transcribed for analysis. All interviews were conducted virtually and led by one interviewer and a note-taker. Questions specific to this study included perceptions of GenAI, how they used GenAI to verify, access, and search for information, and reasons for doing so. Participants were asked to describe their interactions with GenAI, focusing on its specific features and tasks they were performing. Specifically, we asked participants to critically reflect on these experiences with GenAI. The six-phase thematic analysis for qualitative data analysis was conducted (Braun & Clarke, 2006). After reviewing the interview data, the first two authors initiated an initial round of semantic coding to generate inductive codes. They then coded each transcript by highlighting and annotating data items, independently identifying emerging patterns. Any conflicts regarding such themes were

resolved through discussion between the two authors. Through this process, three main themes related to the affordances of GenAI were identified.

Results

Preliminary results reveal that the three main themes include 1) content searching and curation, 2) content generation and ideation, and 3) content enhancement and refinement. Additionally, participants interviewed mostly used ChatGPT, and a few shared they used other GenAI tools (e.g., Grammarly, Gemini, Co-pilot) occasionally.

Allows simultaneous content searching and curation

Participants use GenAI as an alternative to traditional search engines by directly keying in questions they have into GenAI platforms to see what insights they can derive from GenAI's output. Unlike traditional search engines such as Google which presents a list of separate relevant sources in response to a query, GenAI systems can consolidate desired information from different sources and present it through human-like interactions. This makes it easier to use GenAI to find information on a specific topic, eliminating the need to look through various sources. This viewpoint is echoed by various participants and here is an example from a male participant (Electronic and Information Engineering),

> *I just went to search what labs they have for my department, and the kind of research they do. It comes out a lot easier instead of having to scroll through. There's a lot of information on the website.*

Notably, content information sought and consolidated by participants varied widely, spanning topics from makeup to motorcycles and personal details to publicly available information. This diversity highlights how GenAI serves as a resourceful tool for content curation across a wide array of subjects, catering to both casual inquiries and in-depth research. This affordance also aligns with Gen Z's experience of FOMO (Fear of Missing Out), which often drives their information-seeking behaviors as they strive to stay updated.

Encourages content generation for preliminary ideation

Participants use Gen AI to generate ideas and brief pointers, which serve as a starting point upon which they can further build and edit. This is observed in both academic contexts and non-academic contexts in everyday life. A male participant (Economics and Data Science) explained, '*I would prompt ChatGPT that I am applying for [a] scholarship and ask what I should say to the hiring officer.*' Here, the emphasis on content generation is on academics.

Interestingly, participants also indicate that GenAI can be a source of inspiration for other miscellaneous tasks in everyday life. For instance, a female participant (Communication Studies) provided an example, '*to generate a packing list of what I would need to move in [to hostel]. I kind of use that to ensure I don't miss out on anything.*' Another female participant (Psychology) shared, '*when I'm baking or cooking, maybe I would use ChatGPT to search for some recipes.*'

Across various purposes, the generated content was primarily utilized for preliminary ideation and brainstorming where participants asked GenAI to generate ideas on a wide range of topics in everyday life (e.g., recipes and scholarship interviews). Findings indicate GenAI's versatility in enhancing creativity and providing quick access to '*initial*' information aligning with the '*information opportunities*' highlighted by Hirvonen et al. (2023).

Iterates to enhance and refine content

Apart from generating a first draft (ideation), participants used GenAI to gather feedback on a piece of work they had previously created. Thus, participants could refine the original work through multiple iterations. Because all participants were interviewed right before they entered university, they talked about their experiences during university and scholarship applications, which may require essays as part of the process. Here, GenAI was used to improve university application

essays or resumes and this viewpoint was highlighted by many participants. A female participant (Business Administration) informed,

> *I put in a paragraph of everything I want and if I'm not satisfied with the answer, I will refine it like, 'Can you make it more concise.' And I'll do it until I'm happy with the answer.*

GenAI's ability to curate past interactions plays a crucial role in improving the content refining process, particularly by catering to Gen Z's short attention spans. By retaining and referencing previous exchanges, GenAI tools offer more relevant and personalized responses through multiple iterations, reducing the need for users to remember the details and explain their queries repeatedly. Continuity is an important affordance that allows for more efficient enhancement and refinement.

Discussion

This preliminary study identified three types of GenAI affordances (content searching and curation, content generation and ideation, as well as content enhancement and refinement) that facilitate information seeking and shape the information practices of GenZ students. Together, these affordances highlight GenAI's potential to transform traditional information-seeking behaviors by not only providing access to information but also actively shaping the way users interact with, interpret, and apply the information. Indeed, as digital natives, Gen Z is well-positioned to leverage GenAI's affordances, unlocking new '*information opportunities*' surrounding the three themes to create values during information seeking.

The present study's approach to GenAI affordances sheds light on practical use cases and on how GenZ students perceive and utilize the technology to achieve their goals. First, the uniqueness of GenZ's GenAI usage stems from their greater comfort and familiarity with digital technology compared to older generations. Their inherent ease with technology gives them the heightened ability to discover GenAI affordances independently, and quickly adapt these affordances into diverse aspects of their lives, from academic research to everyday problem-solving. Second, the affordance-based approach focuses on identifying how GenAI provides specific opportunities for users to interact with the technology effectively. For GenZ users, this means understanding how AI's features and capabilities enable them to perform various informational activities.

Affordances perceived by GenZ can cause problems in the long run. Firstly, because GenAI affords the generation of original content, there are potential negative implications for human creativity if there is an overreliance on technology for idea generation. This raises the question of whether GenZ students will begin producing uniform answers inspired by AI-generated content, instead of fully developing their creative ideas. Specifically, Google's recent advertisement on their GenAI platform Gemini showcased a father using it to help his daughter write a letter to Olympic athlete Sydney McLaughlin-Levrone, sparking controversy about the role of GenAI in producing creative work (Duffy, 2024). There is a need to discuss where ethical boundaries should be drawn and to guide GenZ students in navigating these considerations.

Secondly, using GenAI for information searching leads to an increased risk of misinformation. Specifically, a phenomenon called hallucinations, where the GenAI model generates false information that does not align with real-world facts, is a major risk for GenZ students (Lee, 2023). The problem is exacerbated by a lack of transparency over how exactly the model arrives at an output or which sources are used, making it difficult to verify AI-generated information. GenAI literacy programs in higher education need to be enhanced so GenZ students are taught to assess the credibility, accuracy, and reliability of GenAI content. In particular, understanding the '*checkworthiness*' of GenAI content, which refers to content that needs verifying before accepting it as true or using it, is an important first step for GenZ students (Wright & Augenstein, 2020).

Conclusion

This study reveals the perceived affordances of GenAI for GenZs, addressing a gap in the current literature on GenAI. The findings underscore the significant extent to which GenAI has been

integrated into the daily lives of GenZ students. Due to the '*technoholic*' nature of GenZs, they are more likely to embrace AI technology in different areas of their life compared to millennials or Generation X (Chan & Lee, 2023).

This study's research and practical contributions should be noted. Using affordances as a framework to examine how GenAI affords resources to GenZs, this study identified affordances that contribute to the literature and raise awareness of how Gen AI can be integrated into formal and informal education settings. More broadly, our findings can help inform educators and policymakers about the development of AI literacy policies for GenZs.

We acknowledged that the relatively small number of participants (n=34) limits the generalizability of this research. Nevertheless, the qualitative results can serve as a helpful starting point on GenZs' information practices surrounding GenAI. Additionally, self-reported data on GenAI use in this study did not capture all aspects of GenAI use, such as frequency, variety of applications, or interactions with different types of AI models. Future work can use a mix of self-reports, usage logs, and analytics to obtain a more comprehensive view of GenAI use.

Acknowledgements

This research is supported by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No: AISG-CV-2023-013).

About the authors

Chei Sian Lee is a Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University in Singapore. Her current research centers on generative AI and digital nudging, aiming to enhance engagement and learning through an information-driven approach. She can be contacted at leecs@ntu.edu.sg.

Li En Tan is a Research Intern at the Wee Kim Wee School of Communication and Information, Nanyang Technological University in Singapore, and is currently an undergraduate student at the University of Cambridge. She can be contacted at lientan080@gmail.com.

Dion H. Goh is a Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. His research interests include social media practices and perceptions, game-based techniques for shaping user perceptions and motivating behaviour, and online information sharing and seeking. He can be contacted at ashlgoh@ntu.edu.sg.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chan, C. K. Y., & Lee, K. K. W. (2023). The Ai Generation Gap: Are Gen Z students more interested in adopting generative ai such as chatgpt in teaching and learning than their gen X and millennial generation teachers? Smart Learning Environments, 10(1), 60. https://doi.org/10.1186/s40561-023-00269-3

Davis, J. L. (2020). How artefacts afford: The power and politics of everyday things. The MIT Press.

Dennen, V. P., & Bagdy, L. M. (2024). Transitions in social media use and social media use for transition: A survey study of the shift from high school to college. Proceedings of the International Conference on Networked Learning, 13. https://doi.org/10.54337/nlc.v13.8546

Duffy, C. (2024, August 2). Google pulls Olympics ad that showed ai writing a little girl's letter for her. CNN. Retrieved August 30, 2024, from https://edition.cnn.com/2024/08/02/tech/google-olympics-ai-ad-artificial-intelligence/index.html

Gibson, J. J. (1986). The ecological approach to visual perception. Hillsdale, NJ: Lawrence Erlbaum Associates.

Giunta, C. (2017). An emerging awareness of generation Z students for Higher Education Professors. Archives of Business Research, 5(4), 90-104. https://doi.org/10.14738/abr.54.2962

Hicks, A. (2022). Negotiating change: Transition as a central concept for information literacy. Journal of Information Science, 48(2), 210-222.

Hirvonen, N., Jylhä, V., Lao, Y., & Larsson, S. (2023). Artificial Intelligence in the information ecosystem: Affordances for everyday information seeking. Journal of the Association for Information Science and Technology, 75(10), 1152-1165.

Howard, H. A., Bochenek, A., Mayhook, Z., Trowbridge, T., & Lux, S. (2023). Student information use during the COVID-19 pandemic. The Journal of Academic Librarianship, 49(3), 102696.

Lee, C. S. (2010). Managing perceived communication failures with affordances of ICTs. Computers in Human Behavior, 26(4), 572–580.

Lee, C.S. & Li, B., & Wu, Q., (2023a). Teenagers and Videoconference Fatigue: A Preliminary Analysis from an Affordance-based Approach. Proceedings of the Association for Information Science and Technology. 60. 635-640. 10.1002/pra2.830

Lee, C.S., Yeo, R.Q.Y., Goh, D.HL., Ang, R.P., Ng, B. (2024). A Transitional Approach to Examine the Influences of Information Practices on Lifelong Learning. Proceedings of the Association for Information Science and Technology. 61. 565-570. 10.1002/pra2.830

Lee, C.S., Yeo, R.Q.Y., Zhang, S., Goh, D.HL., Ang, R.P., Ng, B. (2023b). Gen Z Transitions to University: A Preliminary Study on Everyday Information-Seeking Practices. In: Goh, D.H., Chen, SJ., Tuarob, S. (eds) Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration. ICADL 2023. Lecture Notes in Computer Science, vol 14457. Springer, Singapore.

Lee, M. (2023). A mathematical investigation of hallucination and creativity in GPT models. Mathematics, 11(10), 1–17.

Lim, K.K., & Lee, C.S. (2024). Learning Beyond the Classroom in the AI Era: A Generation Z Perspective. Proceedings of the HCI International 2024 Posters. HCII 2024. Communications in Computer and Information Science Communication, vol 2117. Springer, Cham. https://doi.org/10.1007/978-3-031-61953-3_18

Lloyd, A. (2005). No man (or woman) is an Island: Information literacy, affordances and communities of practice. The Australian Library Journal, 54(3), 230–237.

Medina, J. (2008). Brain Rules. Washington, Seattle: Pear Press

Mosca, J. B., Curtis, K. P., & Savoth, P. G. (2019). New approaches to learning for Generation Z. J. Bus. Divers., 19(3), 66–74.

Mulaudzi, I. C. (2023). Challenges Faced By First-Year University Students: Navigating the Transition to Higher Education. Journal of Education and Human Development, 12(2), 79–87.

Nicholas, A. (2020). Preferred Learning Methods of Generation Z. Faculty and Staff - Articles & Papers. 74. https://digitalcommons.salve.edu/fac_staff_pub/74

Schrock, A. (2015). Communicative Affordances of Mobile Media: Portability, Availability, Locatability, and Multimediality. International Journal of Communication, 9, 18. Retrieved from https://ijoc.org/index.php/ijoc/article/view/3288

Seemiller, C., & Grace, M. (2017). Generation Z: Educating and engaging the next generation of students. About Campus, 22(3), 21–26.

Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a theory of interactive media effects (TIME). In The Handbook of the Psychology of Communication Technology (pp. 47–86). John Wiley & Sons, Ltd.

Taecharungroj, V. (2023). What can ChatGPT do? Analyzing early reactions to the innovative AI chatbot on Twitter. Big Data and Cognitive Computing, 7(1), 35.

Tomaszewski, R. (2023). Library snackables: A study of one-minute library videos. The Journal of Academic Librarianship, 49(2), 102647.

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. Journal of Second Language Writing, 62, 101071.

Wright, D., & Augenstein, I. (2020). Claim check-worthiness detection as positive unlabelled learning. Findings of the Association for Computational Linguistics: EMNLP 2020, 476–488.

Zhitomirsky-Geffet, M., & Blau, M. (2017). Cross-generational analysis of information seeking behavior of smartphone users. Aslib Journal of Information Management, 69(6), 721–739.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

On the robustness of cover version identification models: a study using cover versions from YouTube

Simon Hachmeier and Robert Jäschke

DOI: https://doi.org/10.47989/ir30iConf47077

Abstract

Introduction. Recent advances in cover version identification have shown great success. However, models are usually tested on a fixed set of datasets which are relying on the online cover version database SecondHandSongs. It is unclear how well models perform on cover versions on online video platforms, which might exhibit alterations that are not expected.

Method. We annotate a subset of versions from YouTube sampled by a multi-modal uncertainty sampling approach and evaluate state-of-the-art cover version identification models.

Results. We find that existing models achieve significantly lower ranking performance on our dataset compared to a community dataset. We additionally measure the performance of different types of versions (e.g., instrumental versions) and find several types that are particularly hard to rank. Lastly, we provide a taxonomy of alterations in cover versions on the web.

Conclusions. We found that research in cover version identification shall be less dependent on SecondHandSongs but rather on more diverse datasets.

Introduction

In the context of western popular music, a cover version is a derivative of an original performance of a musical work. Artists perform versions to convey their subjective interpretations of musical works, which is a long-standing practice in musical culture. Usually, different versions of the same work share similar changes of individual notes (melody) or groups of notes (harmony) over time (Yesiler et al., 2021).

The research field of version identification (VI) deals with the automatic detection of cover versions in music collections. Recent approaches in VI aim to encode versions into representations retaining only relevant information in the context of cover versions (Du et al., 2021, 2022, 2023; Hu et al., 2022; Liu et al., 2023; Yesiler et al., 2020a, 2020b). For instance, Abrassart and Doras (Abrassart & Doras, 2022) report that melody, harmony, and lyrics are generally more relevant than rhythm. However, the actual relevance of each characteristic is non-trivial to predict and might strongly vary for different musical pieces. In contrast, characteristics irrelevant in the VI context are usually well agreed upon, such as the tempo or the key/scale.

Online video platforms feature various application scenarios for VI such as copyright infringement detection and music recommendation. Hence, the robustness of methods against noise and variance on the platform is important. One key peculiarity of VI in online videos is the alignment problem. In VI, this was addressed by summarization of musical content along the time axis including pooling mechanisms (Du et al., 2021, 2022; Yesiler et al., 2020a; Yu et al., 2020) and more recently by the matching of smaller chunks of the pairs (Du et al., 2023; Liu et al., 2023). Since YouTube is a collection of videos rather than versions (Except for YouTube's proprietary music streaming service *YouTube Music*), the relationship between videos and versions is an m-to-n relationship. This makes the alignment problem in online videos particularly challenging. For instance, a video might contain multiple versions (e.g., concert recordings) or only chunks of a version (e.g., guitar solo covers or tutorials (Hanson, 2018)). Additionally, videos might include noise such as commentary (e.g. people reacting to music (McDaniel, 2021)). Beside the alignment problem, other challenges might arise for VI in online videos such as the absence of the main melody (e.g. karaoke or instrumental versions (Agrawal & Sureka, 2013; Smith et al., 2017)), low fidelity in amateur recordings and versions occurring only in the background as accompaniment (Martet, 2016).

VI research has made great progress in recent years, mainly measured in metrics from MIREX (https://www.music-ir.org/mirex/wiki/2021:Audio_Cover_Song_Identification) and reported on community datasets like SHS100K-Test (Xu et al., 2018) and Da-Tacos (Yesiler et al., 2019). However, both of these datasets are based on the platform SecondHandSongs (SHS) (https://secondhandsongs.com/) curated by a community of volunteers (https://secondhandsongs.com/page/About) which makes present cover version collections subject to the selection policies of the platform. For example, *web covers* are considered an individual category of versions characterized by being released non-commercially (https://secondhandsongs.com/page/Guidelines/Entities/WebCover). At the same time, they appear to be less relevant for collaborators, since the amount of web covers is usually much lower than for commercially released covers as can be seen for the example ``Enter Sandman'' by Metallica (https://secondhandsongs.com/work/6616). What is more, due to a technical limitation of the application interface of SHS, the created datasets do not actually contain web covers. This poses the question whether VI models trained and evaluated on data from SHS are considering all relevant characteristics of versions and motivates our first research question:

- **RQ1:** do cover version datasets based on the platform SecondHandSongs represent the distributions of cover versions and their characteristics on YouTube?

We assume that there exists a subset of versions with specific characteristics on YouTube which are relevant in the context of VI but not found on the platform SHS: out-of-distribution data. Consequently, recent VI models are neither trained nor evaluated on data with regard to these characteristics. We therefore propose our second research question:

- **RQ2:** which characteristics of versions drive the uncertainty of existing VI models?

In this paper, we aim to explore the success and the challenges of VI on out-of-distribution data. Rather than relying on the cover version collection SecondHandSongs, we leverage the richness of creativity of the YouTube community. Applying a multi-modal uncertainty sampling approach, we identify the most uncertain version candidates. Subsequently, we obtain human annotations by workers on the crowdsourcing platform Mechanical Turk (MTurk). Lastly, two music experts curate a subset of the dataset and provide annotations of uncertainties in the problem context, together with a taxonomy of these.

In summary, the main contributions are:

- we provide a benchmark dataset SHS-YT (https://github.com/progsi/SHS-YT) created with a multi-modal uncertainty sampling approach followed by human annotations. It includes labels on an ordinal scale to reflect the complexity of VI on online video platforms (e.g., videos without musical content and identical audio tracks).
- two experts curate the provided dataset to gather insights into uncertainties in the VI context of online video platforms. We also provide a taxonomy extending an existing one (Yesiler et al., 2021).
- our benchmarks show that even the current state-of-the-art model under-performs on our proposed dataset. Additionally, we identify challenging alterations such as the isolation of single instruments or the vocal track which would be better addressed in the field of query-by-humming. This uncovers potential boundaries of cover version definitions.

Related work

Version identification datasets

VI datasets are composed of versions which are grouped by musical works. During training, VI models are optimized to encode audio representations of versions of the same work as similar and versions of different works as dissimilar. In the evaluation scenario, each version represents a query at a time and the remaining versions are ranked based on the musical similarity computed by the VI model. The resulting N rankings for a dataset with N versions serve as the input to compute retrieval metrics such as the mean average precision (MAP).

We provide an overview of the most popular datasets in VI which are used for benchmarking compared to the datasets used in this paper in Table 1. Recent VI approaches (Du et al., 2021, 2022, 2023; Hu et al., 2022; Liu et al., 2023; Yesiler et al., 2020a, 2020b) achieve MAP scores up to 0.96 on YouTubeCovers (Silva et al., 2015) and Covers80 (Ellis, 2011). The results on the larger datasets SHS100K-Test (Xu et al., 2018) and the Da-Tacos benchmark subset (Yesiler et al., 2019) are lower; for instance, CoverHunter (Liu et al., 2023) achieves the highest MAP but still does not surpass 0.90.

Dataset	Works	Versions
Covers80	80	160
Da-Tacos	1,000	15,000
Discogs-VI-YT-Test	9,878	116,547
SHS100K-Test	1,692	10,547
YouTubeCovers	50	350
SHS-SEED	100	2,404
SHS-YT	100	900
SHS-YT+2Q	100	1,092
SHS-YT+AllQ	100	3,289

Table 1. Popular VI benchmark datasets and the seed dataset and our annotated datasets in bold text

A commonality of all of these datasets but Covers80, is their utilization of SHS as a data source. The same accounts for the respective training sets of VI models: SHS100K-Train and the training subset of Da-Tacos (Yesiler et al., 2020a) which were used to train the recent VI models. Consequently, versions in the dataset can be found on YouTube, but are only included if these are manually collected by the SHS community. The question remains whether the distribution of variance of versions on YouTube is appropriately represented in existing benchmark datasets. A newer dataset, namely Discogs-VI-YT, is based on Discogs (https://www.discogs.com/) rather than SHS. It is the currently biggest dataset in VI. Since it is rather new, there are currently no benchmarks of the state-of-the-art VI models.

Music on YouTube

Various studies address the richness and diversity of versions on YouTube and corresponding classes. In Table 2 we distinguish between 4 classes of versions and provide some examples found in existing research. Liikkanen and Salovaara (2015) state that music is the most popular content type on YouTube. The results were derived from data about YouTube search trends, the most popular videos, and channels. The authors established twelve subclasses of versions segmented into three main classes: official (uploaded by copyright owners), user-appropriated (uploaded by fans) and derivative (e.g., cover versions). While the first two classes are expected to contain highly similar audio, the third class rather relies on music fans and hobby musicians. It includes stronger changes in musical characteristics; for instance, covers on instruments, parodies, or remixes.

The category of user-appropriated versions is also discussed by Martet (2016). The author also includes a new perspective on versions, including videos which contain versions rather as an accompaniment (e.g., for movie trailers).

Dataset	Works
Official	Official Music Video Professional Live Video
User-Appropriated	Lyric Video Slideshow
Cover	Guitar Cover Paraody Karaoke Version
Other	Tutorial Reaction Video

Table 2. Classes and examples of versions on YouTube.

From an application-driven perspective, studies have implemented pipelines to cope with copyright infringement detection (Agrawal & Sureka, 2013) and music retrieval (B. Li & Kumar, 2019; Smith et al., 2017) on YouTube. Smith et al. (2017) propose an approach processing audio, text, and video features to predict a version class. Similar to Airoldi et al. (2016) as well as Liikkanen and Salovaara (2015), the authors established classes like remixes and tutorials beside official music videos and live performances. Another approach to model classes of versions on YouTube derived clusters of categories of versions by a network analysis (Airoldi et al., 2016). The results emerged clusters corresponding to musical genres and situational contexts (Eg. covers and tutorials).

While the classes of versions in all of these studies might be relevant for VI research, their consideration in the field is rather limited. Yesiler et al. (2021) construct a taxonomy where they also mention some classes and the corresponding alterations of musical characteristics. To best of our knowledge, no existing benchmarks of VI models investigated the impact of the mentioned alterations on model robustness.

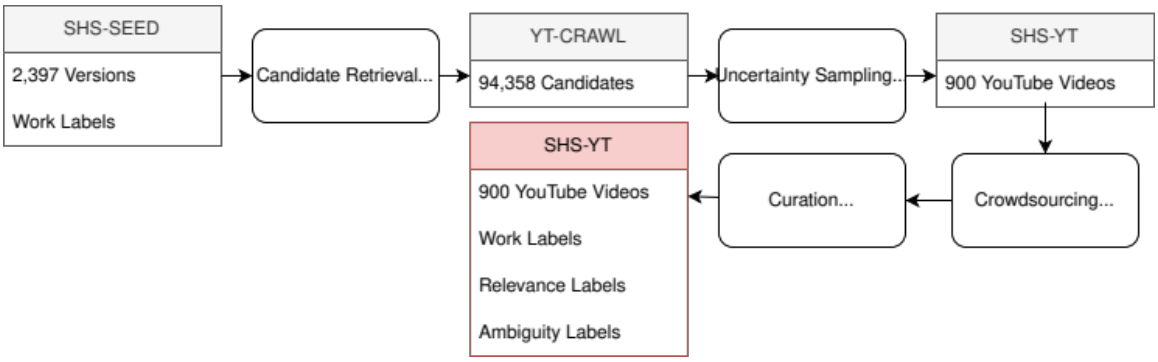


Figure 1. Dataset creation

Data creation

We here describe the steps of the creation process of the dataset SHS-YT as shown in Figure 1.

We aim to evaluate the performance of state-of-the-art VI models on out-of-distribution data. We select YouTube as a rich source for a diverse set of versions, since there are no constraints for uploaders as opposed to the policies on SHS.

To cover a representative subset of western popular music, we select the widely used SHS100K-Test as a seed dataset composed of works of western popular music. In particular, we choose the first 100 works from its test subset (https://github.com/NovaFrost/SHS100K/blob/master/SHS100K-TEST). These works are represented by 2,859 versions of which we successfully retrieved 2,397. We denote this dataset with 2,015 unique performers as *SHS-SEED*.

Candidate retrieval

The goal of the candidate retrieval step is to obtain a set of candidate versions to be included in our dataset. We apply an approach by Hachmeier et al. (2022) to formulate multiple text queries per work in SHS-SEED. We utilize the strings for performer and title of the first version of each work to formulate queries and additionally formulate new queries using YouTube search suggestions (The list of queries per work can be found in our repository). On average, we formulate 44 text queries per work resulting in 4,365 text queries. We retrieve metadata for the top 50 videos per query (https://pypi.org/project/youtube-search-python) and drop videos with a length of 10 minutes or more. We denote the resulting collection of 94,358 videos as YT-CRAWL. We download the audio files for all videos with a sampling rate of 22,050 Hz(https://github.com/yt-dlp) and extract CREMA features which we use in the next step (https://github.com/bmcfee/crema).

Uncertainty sampling

In the uncertainty sampling step, we aim to reduce the number of versions to a smaller subset because of two reasons: Firstly, we are limited in annotation capacity. Secondly, we aim to focus on out-of-distribution data and want to focus on versions with characteristics not common to be found on SHS. We leverage the modalities of audio (CREMA features) and text (YouTube metadata). For both domains, we use models based on deep learning as proxies. In theory, only the audio information is necessary to determine whether two versions are associated with each other. However, we use the text-based proxy to systematically find candidates where the VI proxy over- or underestimates the musical similarity.

Modality proxies

We use the pre-trained model Re-MOVE (Yesiler et al., 2020b) as a proxy in the audio/music domain which is one of the best approaches for VI at the time of dataset creation. The model processes CREMA features, which represent harmonic and melodic progressions, and encodes these into 256-dimensional embeddings. The Cosine similarity of a pair of embeddings represents their musical similarity. For the text domain, we use the entity matching model Ditto (Y. Li et al., 2020). The model is based on BERT (Devlin et al., 2018) and encodes pairs of textual entities into BERT embeddings and predicts a binary matching confidence. From the SHS100K-Train dataset we create a train, validation, and test set with a ratio of 3:1:1 as proposed by Y. Li et al. (2020) with each containing positive and negative pairs of YouTube videos in a 1:4 ratio. We gather the negative pairs by randomly sampling videos from another randomly selected work. We use all of the proposed data augmentation techniques and the best performing language model (RoBERTa) as described by Y. Li et al. (2020). We apply the best model checkpoint evaluated on the test set after 50~epochs for our matching task.

Since the inclusion of YouTube descriptions yielded inferior results (F1 score of 0.27 against 0.95) we solely process YouTube titles and channel names.

Similarity and matching confidence aggregation

For each candidate in YT-CRAWL we compute a similarity and matching confidence with the proxy models. Since each of the works is represented by multiple versions in SHS-SEED (24 on average) we must aggregate the pairwise similarities and model confidences. For a work i, a set of query versions from SHS-SEED Q_i and a candidate version c_{ij} from YT-CRAWL, we compute the musical similarity $S_m(c_{ij})$ as the arithmetic mean of the Cosine similarities of the Re-MOVE outputs of all

pairs (c_{ij}, q) for $q \in Q_i$. In a preliminary experiment on the validation dataset of SHS100K we validated the aggregation by the arithmetic mean as opposed to aggregation by maximum. We further compute the textual similarity $S_t(c_{ij})$ for the same pairs as the maximum matching confidence based on Ditto. We motivate this because candidates with non-matching metadata among the queries shall not have impact on the matching decision as long as at least one candidate in Q_i matches. This is especially of relevance in cases with translated version titles. For instance, the version title 'Tiempo de Verano' (Spanish for 'Time of the Summer' or 'Summertime') which is potentially a substring within a YouTube title might match the version title 'Summertime' with rather low confidence. Based on the aggregated values $S_m(c_{ij})$ and $S_t(c_{ij})$ for all candidates in YT-CRAWL, we conduct uncertainty sampling with two approaches: disagreement sampling and mutual uncertainty.

Disagreement sampling

We establish two disagreement groups: *DisagrAudio* denotes the candidates where the musical similarity high in contrast to the textual similarity and *DisagrText* represents the contrary case. We measure the disagreement as the absolute difference as shown in Table 3 and select the three candidates with the highest disagreement for both disagreement groups per work.

Group	Ranking Function
DisagrAudio	$S_m(c_{ij}) - S_t(c_{ij}) if S_m(c_{ij}) > S_t(c_{ij})$
DisagrText	$S_t(c_{ij}) - S_m(c_{ij}) if S_t(c_{ij}) > S_m(c_{ij})$
DisagrUnc	$-\left\|\left(S(c_{ij}), S^*(C_i)\right)\right\|_2$

Table 3. Uncertainty groups and their constraints. We sample the top three results returned by each ranking function.

Mutual uncertainty

We denote the mutual uncertainty group by *DisagrUnc* containing the top three candidates with the highest mutual uncertainty. Works with less than three candidates for DisagrAudio are filled with samples from this group as well. As shown in Table 3, we compute the mutual uncertainty as the negative Euclidean distance between the two-dimensional vector $S(c_{ij}) = \left[S_m(c_{ij}), S_t(c_{ij})\right]^T$ and the vector $S^*(C_i)$, representing the center of uncertainty based on all candidates for the work C_i, defined as follows:

$$(1) \quad S^*(C_i) = \begin{matrix} S_m^*(C_i) \\ S_t^*(C_i) \end{matrix}$$

with

$$(2) \quad S_\theta^*(C_i) = \frac{1}{2}\left(S_\theta^{min}(C_i) + S_\theta^{max}(C_i)\right)$$

where $\theta \in \{m, t\}$ and $S_\theta^{min}(C_i)$ and $S_\theta^{max}(C_i)$ return the minimum and maximum of the Cosine similarities or matching confidences for all the candidates in C_i, respectively. In the following, we describe our annotation process of the resulting nine candidates per work.

Annotation

We impose an ordinal scale of classes and obtain annotations by workers from Amazon's Mechanical Turk (MTurk) and inhouse-experts.

Relevance classes

Prior VI datasets solely consider the membership of a version to a work (binary label). Hence, each item in the dataset is expected to contain music. Further, the versions in the dataset of the same work are expected to be different regarding different aspects such as tempo or timbre (Yesiler et al., 2021). Both is not guaranteed when dealing with our retrieved candidates from YouTube, since videos are not even guaranteed to contain music. We construct four classes on an ordinal scale with respect to the relevance to the query version:

- *NoMusic*: candidate version does not contain music and is not relevant.
- *NonVersion:* candidate version does not contain music and is not relevant.
- *Version*: candidate version is derived of the same work as the query versions and therefore relevant.
- *Match*: candidate version includes (parts of) the exact same audio as the original they are derived of (*user-appropriated* videos). The version is relevant.

We represent each work i by a query version which is a random version from SHS-SEED. The goal of the annotation step is to gather annotations about the relevance between i and the candidate in the set. We denote the resulting set of 900 annotated versions as SHS-YT.

Crowdsourcing

We publish one human intelligence tasks (HITs) on MTurk per work with instructions and examples as shown in Figure 2. Each contains the query version, the nine candidates and a quality check candidate with a known answer *Version* or *NonVersion* based on the works and versions in SHS-SEED. To simplify the task, we specifically instruct that excerpts are sufficient (e.g., a medley is a *Version* if it contains an excerpt of the query).

The interface and manual presented to the workers can be found in our published dataset.

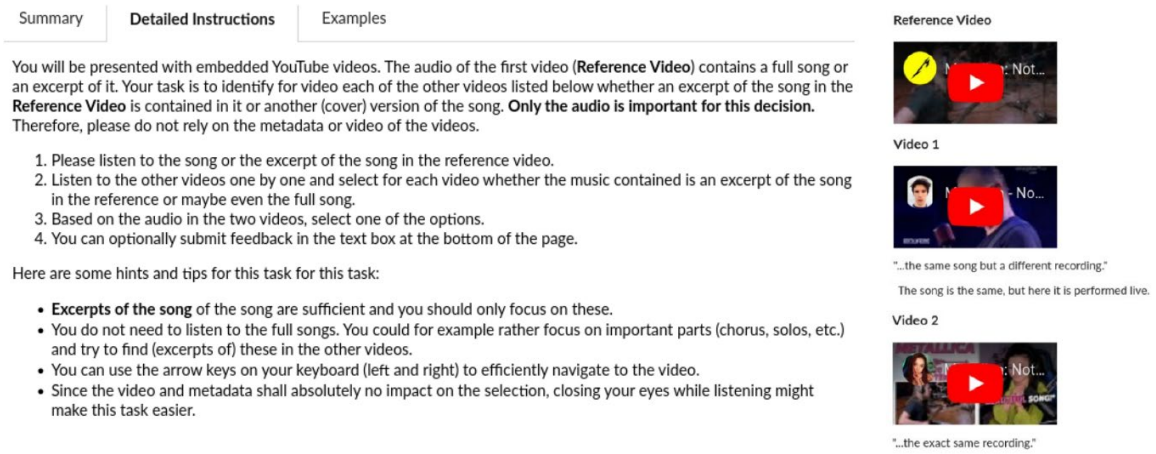


Figure 2. Our instructions and examples to workers as presented on MTurk. Please note that the examples of the right are cropped to fit

We measured the average time effort per annotation pair at 90~seconds and thus expect 15 minutes per HIT. We pay a reward of 3.2 US dollars per HIT corresponding to our domestic minimum wage, compensating our estimated time effort in consideration of the average currency exchange rate between our currency and the US dollar at annotation time.

We collect assignments by up to five workers per HIT. Following best practices to achieve annotation quality (Ghosh et al., 2019; Matherly, 2018; Mellis & Bickel, 2020; Peer et al., 2013) we only permit workers with more than 100 approved HITs and approval rate above or equal to 99%. We

reject assignments where workers fail the quality check or complete the assignments in less than ten seconds. In some cases, we accept assignments with failed quality checks due to proper justifications by workers. We do not include these assignments in our dataset. The final worker labels are obtained by majority voting: minimum three equal labels determine the decision for the label. Candidates which remain without a final label due to high variance in label responses are curated by the experts in the next step.

Curation

We employ two music experts for curation of the annotated dataset (The two experts have 15 years of musical experience on harmonic instruments.). The experts' task is to check the relevance labels of the workers for correctness, decide for a relevance label in undecided cases and to annotate the most prominent reason which makes a candidate more difficult to annotate (uncertainty class). In cases of uncertainty, the experts discuss decisions. Ultimately, experts and authors agreed upon including boundary cases as versions as well (e.g., remixes).

The first expert curates candidates labeled with *NoMusic* and 167 candidates with failed majority votes due to ties or shortage of worker assignments (because of failed assignment quality checks). Based on the collected reasons for uncertainties by the first expert, we formulate uncertainty classes and distinguish between uncertainties related to the version itself (e.g., *Song*: *Instrumental*) and uncertainties related to the Version in context of its occurrence in an online video (e.g., *Video*: *With Non-Music*). Some uncertainty classes just apply to one relationship class, for example, *Song*: *Same Artist* only applies if the candidate is a *NonVersion*. We provide a full documentation in our published repository.

The second expert utilizes the uncertainty classes directly and curates all candidates labeled with Version and the 96 most similar candidates labeled with *NonVersion* (Measured in mean Cosine similarity per benchmarking model as explained in the previous section) for error analysis. New uncertainty classes are collected and iteratively formulated, resulting in a total of 19 ambiguity classes. Based on these classes derived of observed examples, we construct a taxonomy of alterations.

Dataset analysis

Overview

We present the distributions of numerical YouTube attributes of SHS-YT in Figure 3. We observe a strong peak in duration around 3.5 minutes and in uploading dates between 2020 and 2022.

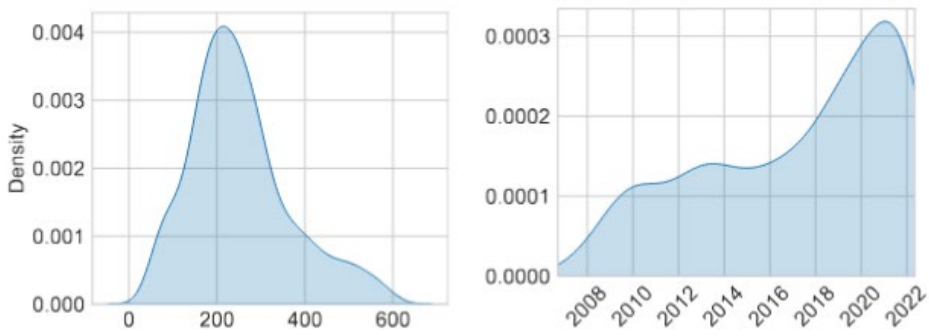


Figure 3. Gaussian kernel density estimates for duration (left) and uploading date (right) of the videos in the SHS-YT dataset. The bandwidth parameter is estimated by Scott's method

In Table 4 we show counts per annotation class and sampling group. The dataset mostly contains versions of other works than their respective query versions but also 197 versions of the same

works. *NoMusic* versions mostly occurred in the DisagrText group, which is expected, since a modeled musical similarity by Re-MOVE is rather unlikely with the absence of actual music. Similarly, the only 4 *Match* versions only occur in the DisagrAudio sampling group. SHS-YT contains 5 versions which are also contained in Da-Tacos; all are labeled with *NonVersion*. Regarding SHS100K, SHS-YT contains 5 versions from the test subset but from other works then in SHS-SEED, 2 from the validation subset and 13 candidates from the training subset. Hence, all of these candidates but one is annotated as *NonVersion*.

	Match	Version	NonVersion	NoMusic	Σ
DisagrAudio	4	89	200	0	293
DisagrText	0	82	142	76	300
DisagrUnc	0	26	280	1	307
Σ	4	197	622	77	900

Table 4. Uncertainty groups and their constraints. We sample the top three results returned by each ranking function

In Figure 4 we show the relative amounts of uncertainty classes excluding placeholder for 104 non-ambiguous versions according to the experts.

Non-musical content is the most represented uncertainty for *Versions* with 14% (n=77), followed by vocal-only. For *NonVersions*, the most frequent uncertainty is musical similarity between versions (*Song: Similar*) at 12%, followed by *NonVersions* from the same artist as the query version with 11%.

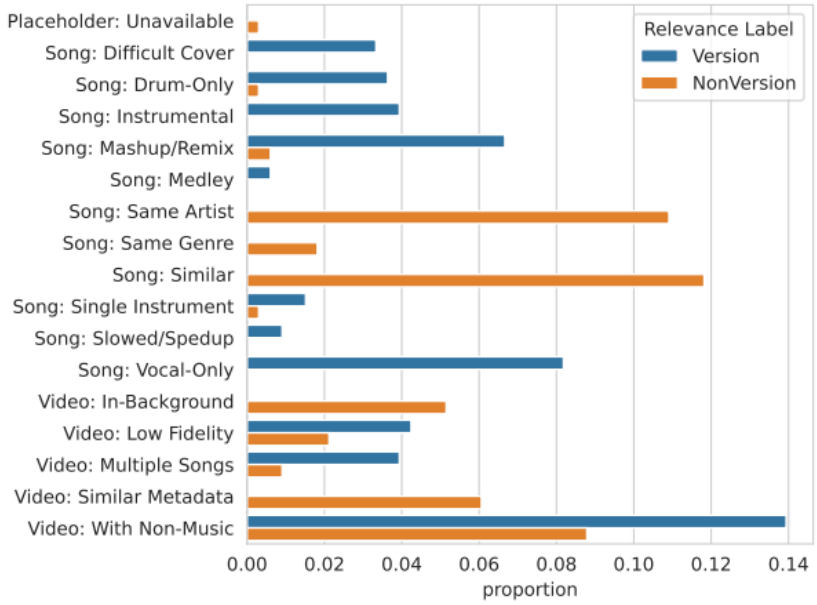


Figure 4. Relative proportion of uncertainty class annotated.

Annotation quality

Comparing the aggregated worker labels with expert labels for our 513 curated versions results in a Kendall's Tau of 0.81, indicating a strong positive association. However, the agreement among workers measured in Krippendorff's Alpha is just 0.43. The moderate level of inter-rater agreement might be partly due to the similarity of the VI task to the audio music similarity task which generally is associated with limited agreement as discussed in previous studies (Daikoku et al., 2020; Flexer et al., 2021; Flexer & Grill, 2016; Flexer & Lallai, 2019; Jones et al., 2007). Looking at the annotated uncertainty classes for candidates that are falsely labeled according to the expert (n=84) or which did not achieve a majority vote (n=167) uncovers some potential issues of workers. Especially versions which include non-musical and musical content seem to confuse workers (n=51). We found examples from 'The Voice' (https://youtu.be/aii62acsp_E) and a movie scene from 'Cocktail' (https://youtu.be/EFuBvEt84OI).

Benchmark

In this section, we conduct a benchmark on our proposed dataset with the goal to gather insights about the VI performance on out-of-distribution data. Since VI is a matching problem, we require relevant versions for all works in the dataset which is not the case for SHS-YT. Hence, we include versions from SHS-SEED. We construct two benchmark datasets derived of SHS-YT which we also show in Table 1. For both datasets, we exclude the versions which are included in the training and validation datasets of SHS100K:

- *SHS-YT+2Q*: SHS-YT with the query versions used for human annotation and one additional work from SHS-SEED. We select the version with the lowest version identifier which either is the original version or at least an earlier version derived of the original. This dataset includes minimum two relevant versions per work. In this dataset of 1,092 versions, our annotated versions account for around 82%. The 312 versions labeled as irrelevant (*NonVersion* and *NoMusic*) account for 29% of the dataset.
- *SHS-YT+AllQ*: Our proposed dataset with the all versions from SHS-SEED resulting in 3,289 versions. Here, our annotated versions account for around 27% of all versions. The 312 versions labeled as irrelevant account for 9% of the dataset.

Beside the modality proxies as described before, we evaluate two other VI models and a fuzzy matching baseline. CQTNet (Yu et al., 2020) is a VI model consisting of mainly convolutional neural networks. It processes constant-Q transform spectrograms (CQTs). The current state-of-the-art model is CoverHunter (Liu et al., 2023) which also processes CQTs but includes a conformer backbone (Gulati et al., 2020) and an attention mechanism (Okabe et al., 2018). The model is trained with a coarse-to-fine training scheme to address the alignment problem.

Both models are trained and validated on SHS100K. We use the pre-trained models provided by the authors. In contrast to the models we benchmark, the approach by Abrassart and Doras (Abrassart & Doras, 2022), LyraCNet (Hu et al., 2022) and the ByteCover models (Du et al., 2021, 2022, 2023) are not publicly available (https://github.com/Orfium/bytecover).

Overall performance

First, we evaluate the performance of models like in traditional VI research and consider only the binary label (relevant or not).

We report two evaluation metrics suggested by MIREX: (https://www.music-ir.org/mirex/wiki/2021:Audio_Cover_Song_Identification) mean average precision (MAP) and mean rank of the first relevant item (MR1). Since precision for the first 10 items is not a fair metric when having works with less than 10 relevant items, we omit this metric in our evaluation.

		SHS-YT+2Q		SHS-YT+AllQ		SHS100K-Test		Da-Tacos	
	Model	MAP	MR1	MAP	MR1	MAP	MR1	MAP	MR1
Audio	CoverHunter	0.52	44.5	0.83	8.1	0.86	11.9	0.85	12.2
	CQTNet	0.50	35.8	0.72	12.4	0.66	54.9	0.74	10.7
	Re-MOVE	0.40	86.9	0.56	18.5	0.53	38.0	0.52	38.0
Text	Ditto	0.39	73.78	0.78	18.5	-	-	-	-
	Fuzzy	0.24	101.3	0.46	14.3	-	-	-	-

Table 5. Benchmark results of VI models, the entity resolution model Ditto and Fuzzy which is the token set ratio from rapidfuzz (Bachmann, 2021)

In Table 5 we report the respective results on our benchmark datasets, SHS100K-Test and Da-Tacos. Please note that we exclude Discogs-VI-YT (Araz et al., 2024). since it was published after our experiments. Furthermore, it has to be noted that both of the evaluation metrics are sensitive to the dataset size which is not negligible (see Table 1). However, smaller dataset sizes usually promote a higher MAP and even though SHS-YT+2Q is a smaller dataset than the others, we observe a rather strong performance drop in MAP between -34% (CoverHunter) and -13% (Re-MOVE). The performance drop is less apparent for CoverHunter at SHS-YT+AllQ and the performance even increases compared to SHS100K-Test for the other VI models. While this is likely due to the larger number of versions from SHS-SEED we further look into the pairwise Cosine similarities for different pairwise relationships in the following section.

A closing remark on the overall evaluation is the potential influence of sampling bias to the performance of Re-MOVE and Ditto since these models are used as modality proxies in dataset creation.

Distributions of Cosine similarities

To support a more well-grounded verdict about the difference of distributions of versions in SHS-YT to versions on SHS and hence in datasets like SHS100K and Da-Tacos, we investigate the Cosine similarities of pairs of versions. A version from SHS-SEED can be considered a baseline version (*SHS-Version*). Our RQ1 aims to uncover whether existing VI models treat two SHS-Versions of the same work as more similar then an *SHS-Version* compared to a version from SHS-YT of the same work (*YT-Version*). Similarly, the question arises whether *NonVersions* from SHS-YT (YT-NonVersion) are more similar than other NonVersions from SHS-SEED (*SHS-NonVersion*): the former are versions in the same YouTube result sets (e.g. of the same artists) and the latter are random other versions.

In Table 6 we show statistics about the respective Cosine similarity distributions of SHS-Versions compared to other types of versions based on the relevance class. We observe that the similarities among SHS-Versions is significantly lower than their similarity to *YT-Versions*. Also, similarities of SHS-Versions of different works are significantly lower than their similarities to *YT-NonVersions*; but the corresponding effect size is lower. Both of these observations are likely a reason for less consistent rankings based on the tested VI models and hence the lower MAP scores observed in the previous section. Additionally, these insights substantiate an answer to RQ1 that in fact there exist different distributions of versions on SHS and YouTube.

Relevance Class	CoverHunter	CQTNet	Re-MOVE	Support
SHS-Version	0.88 ± 0.07	0.61 ± 0.13	0.62 ± 0.16	96,502
YT-Match	0.87 ± 0.08	0.61 ± 0.17	0.66 ± 0.19	44
YT-Version	**0.80 ± 0.10**	**0.48 ± 0.19**	**0.45 ± 0.24**	5,021
SHS-NonVersion	0.68 ± 0.04	0.33 ± 0.08	0.36 ± 0.09	5,637,128
YT-NonVersion	**0.72 ± 0.05**	**0.37 ± 0.09**	**0.41 ± 0.14**	14,305
YT-NoMusic	0.68 ± 0.07	0.23 ± 0.05	0.22 ± 0.05	1,810

Table 6. Arithmetic means and standard deviations of Cosine similarities between the *SHS-Versions* and the respective other versions. The prefix YT- indicates that the version is from SHS-YT and SHS- indicates that it is from SHS-SEED. Bold formatting indicates that means are statistically significant different measured with the Two-Sample-t-Test at $p < 0.01$

	Uncertainty Class	CoverHunter	CQTNet	Re-MOVE	Support
	SHS-Version	0.88 ± 0.07	0.61 ± 0.13	0.62 ± 0.16	96,502
YT-Version	*Version: Difficult Cover*	**0.82 ± 0.11**	**0.55 ± 0.17**	**0.55 ± 0.20**	293
	Version: Drum-Only	**0.72 ± 0.05**	**0.28 ± 0.07**	**0.23 ± 0.06**	321
	Version: Instrumental	**0.68 ± 0.12**	**0.38 ± 0.24**	**0.38 ± 0.28**	364
	Version: Mashup/Remix	**0.76 ± 0.07**	**0.44 ± 0.12**	**0.41 ± 0.21**	518
	Version: Medley	**0.72 ± 0.03**	**0.32 ± 0.09**	**0.25 ± 0.06**	86
	Version: Single Instrument	**0.80 ± 0.05**	0.68 ± 0.13	**0.46 ± 0.10**	195
	Version: Slowed/Sped-up	0.87 ± 0.05	**0.54 ± 0.14**	**0.43 ± 0.24**	63
	Version: Vocal-Only	**0.77 ± 0.04**	**0.38 ± 0.09**	**0.23 ± 0.07**	718
	Video: Low Fidelity	0.86 ± 0.09	**0.57 ± 0.16**	**0.49 ± 0.29**	292
	Video: Multiple Versions	**0.79 ± 0.09**	**0.49 ± 0.17**	**0.52 ± 0.21**	343
	Video: With Non-Music	**0.81 ± 0.10**	**0.48 ± 0.18**	**0.50 ± 0.23**	1,027
	SHS-NonVersion	0.68 ± 0.04	0.33 ± 0.08	0.36 ± 0.09	5,637,128
YT-NonVersion	*Version: Mashup/Remix*	**0.78 ± 0.04**	**0.42 ± 0.07**	0.33 ± 0.14	53
	Version: Same Artist	**0.76 ± 0.04**	**0.45 ± 0.08**	**0.51 ± 0.11**	862
	Version: Same Genre	**0.75 ± 0.05**	**0.40 ± 0.09**	**0.53 ± 0.13**	169
	Version: Similar Version	**0.76 ± 0.06**	**0.45 ± 0.08**	**0.51 ± 0.11**	1,069
	Video: Multiple Versions	**0.71 ± 0.05**	**0.39 ± 0.08**	**0.40 ± 0.14**	102

Table 7. Arithmetic mean and standard deviation of Cosine similarities between versions in SHS-SEED and a version from SHS-YT grouped by the uncertainty class. Bold formatting indicates that means are statistically significant different measured with the Two-Sample-t-Test at $p < 0.01$

To address RQ2, we investigate the differences of Cosine similarities for subsets of relevance classes grouped by their corresponding uncertainty classes in Table 7. Our imposed ordinal relevance classes also allow for analysis of similarities when dealing with highly similar versions (*YT-Match*) and versions without music (*YT-NoMusic*). Interestingly, the similarities are neither significantly higher nor lower than the similarities to other *SHS-Versions*. Regarding *NoMusic* versions, we can also see rather high similarities which indicates a lack of robustness of VI models. Almost all the

YT-Versions are significantly less similar compared to SHS-Versions ($p < 0.01$). The most challenging classes for all the models appear to be drum-only versions, instrumental versions, and medleys. While the latter is rather attributed to an alignment problem, the other two are most likely affected by the absence of the main melody and partly the harmony. Vocal-only versions which most likely only contain the main melody, appear to be hard for CQTNet and Re-MOVE and less so for CoverHunter. Difficulties for VI models for YT-NonVersions appear to arise due to versions being of the same artist, genre or just because they are similar by chance (*Version: Similar Version* and *Version: Mashup/Remix*).

In Figure 5, we further investigate the mean similarities by CoverHunter of different relevant versions. Apparently, the difficulty of drum-only versions is validated. We can also see that versions referring to multiple versions or including non-music noise impact the similarity. In the next section, we provide some examples for versions on YouTube which appear to be very challenging.

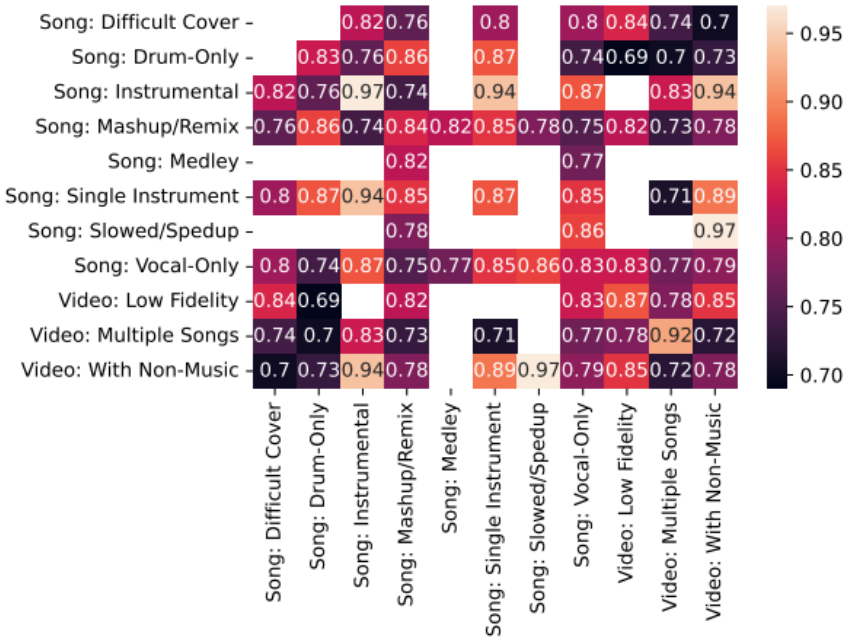


Figure 5. Mean Cosine similarities of CoverHunter embeddings between YT-Versions per respective uncertainty class

Error analysis

We examine the reasons for uncertainties more profoundly. First, we look at versions labelled *Non-Version* which are more similar than random other versions. We found that songs of the same genres are generally more similar, for instance bossa nova and blues. In theory VI models are not optimized to model genres per se. However, musical characteristics such as chord progressions (e.g., blues scheme) or rhythm (e.g., bossa nova beat) seem to be hard to disentangle from VI representations. Similarly, musical characteristics appear to correlate for versions of the same artists (e.g., Lady Gaga, Backstreet Boys, AC/DC). However, in some cases versions appear to be similar simply by similar chord progressions (e.g., 'Ultraviolence' (https://youtu.be/ZFWC4SiZBao) and 'Radioactive' (https://youtu.be/E5sVhFnrlTw)). Interestingly, we also found *NoMusic* versions with high similarity to Versions according to CoverHunter (https://youtu.be/PG6iJmbnOTY and https://youtu.be/svQD6mGDPXc). We assume that this is due to the matching of mute or low energy sections in these versions with mute parts of SHS-Versions after the alignment module.

Investigating some YT-Versions which appear to be difficult to detect, we found that vocal-only can refer either to versions with isolated voice stem by sound source separation (https://youtu.be/cixhJpyTWko) but also self-recorded vocal-only versions (https://youtu.be/24AKYyNusvs).

Discussion and implications

We summarize the findings gathered by our created dataset SHS-YT and the respective benchmarks. Regarding RQ1 we in fact confirmed a significant difference between some of the versions on YouTube and the ones included in SHS-based datasets. Based on our ordinal relevance labels, we derive that the difficulty especially arises due to relevant versions which are difficult to detect (false negatives) rather than irrelevant versions (false positives). However, some aspects such as similarity of songs within genres, of the same artists or with similar chord progressions seem to impact the overestimated similarity.

Looking at our dataset with annotated uncertainty classes reveals that the drum-only videos are rather challenging as well as instrumental versions. While the former do not include melody and harmony, these cases can be denoted as boundary cases. This raises the question about how a cover version is defined which is a question to be asked in musicology and maybe even of philosophical nature. Beside these rather song-specific uncertainty classes, there are also observable difficulties arising due to the alignment problem. While this is a general problem in VI, extreme cases such as medleys, multiple versions in a video and videos with versions and non-musical content still appear to be difficult for existing models.

To improve VI models in the future one solution is to rely on broader datasets in terms of data sources. For instance, by utilizing YouTube metadata to train weakly-supervised models. However, we propose another solution based on our observations. In Figure 6, we propose our taxonomy of cover versions in online videos. In the context of VI, musical characteristics which are discussed by Yesiler et al. (2021) (*Song* node) are one key component to model cover version relationships. Researchers are well aware about the importance of alterations in these characteristics and address them by augmentation techniques such as pitch-variations, tempo-variations. In the context of VI on YouTube (*Video* node), there are additional challenges which arise due to the context of online videos. Our observations provided examples for versions with low-fidelity and versions which occur in the background with foreground noise. We believe that both of these alterations can be well addressed by incorporating audio fingerprinting and noise mixing. We also found that isolated stems (e.g., drum-only, vocal-only versions) are particularly challenging. This is a problem which points to the related music information retrieval task of query-by-humming, where audio representations rely on single stems (usually the singing voice). In VI, an integration of sound source separation in augmentation techniques could further benefit model performance. Alternatively, rather than extracting the features in an end-to-end fashion using CQT spectrograms, one could extract features for melody, harmony, and rhythm separately like Abrassart & Doras (2022).

Lastly, the alignment problem which we have mentioned appears to be particularly strong on online video platforms. Not only can a version be represented only by a section (*Chunked*), but also along with multiple other versions or non-music noise.

The application of sliding time windows, possibly with different sizes followed by a maximum aggregation can address this problem. However, this might in turn increase the risk for false negatives and the computational load. We propose that also the synthesis of data by concatenation of different versions and non-musical noise such as commentary can help to make VI models more robust for these cases.

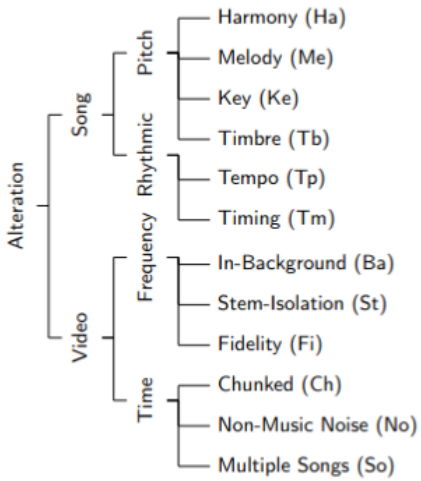


Figure 6. Taxonomy of cover versions in online videos

Conclusion and limitations

In this paper we proposed SHS-YT, a new benchmark dataset for VI. Created with a multi-modal uncertainty sampling approach and annotated by workers and experts including uncertainty classes, this dataset provided novel insights in the robustness of VI models. Lastly, we want to highlight some limitations of our study.

To the best of our knowledge, this is the first study which evaluates VI approaches with regard to different alterations among versions focusing on the most prominent uncertainty. Nevertheless, these classes might be partly subjective and cannot be fully isolated by other effects which might occur for certain pairs of versions. Due to the peculiarity of YouTube of being a dynamic online video platform, we cannot guarantee the presence of our videos on the platform in the future. In our repository, we provide all the URLs investigated. Due to copyright issues, we cannot provide the raw audio but only the extracted CQT and CREMA features. This paper focused on cover versions in the context western popular music. We are well aware that other genres might incorporate other characteristics which make this study less applicable. In future studies, the consideration of other genres with different characteristics could improve to gather an even broader overview of musical reinterpretations.

Acknowledgements

We thank the music experts for supporting the annotation processes.

About the authors

Simon Hachmeier is a Ph.D. student at the Berlin School of Library and Information Science at the Humboldt-Universität zu Berlin. He received his M.Sc. in Information Systems from the University of Innsbruck. Simon's research interest is cover version identification on the web. He can be contacted at simon.hachmeier@hu-berlin.de

Robert Jäschke is a Professor at the Berlin School of Library and Information Science at the Humboldt-Universität zu Berlin. He received his Ph.D. in Computer Science from the University of Kassel. Robert's research interests are web science and digital humanities. He can be contacted at robert.jaeschke@hu-berlin.de

References

Abrassart, M., & Doras, G. (2022). And what if two musical versions don't share melody, harmony, rhythm, or lyrics? International Society for Music Information Retrieval Conference.

Agrawal, S., & Sureka, A. (2013). Copyright Infringement Detection of Music Videos on YouTube by Mining Video and Uploader Meta-data. In V. Bhatnagar & S. Srinivasa (Eds.), Big Data Analytics (pp. 48–67). Springer International Publishing. https://doi.org/10.1007/978-3-319-03689-2_4

Airoldi, M., Beraldo, D., & Gandini, A. (2016). Follow the algorithm: An exploratory investigation of music on YouTube. Poetics, 57, 1–13. https://doi.org/10.1016/j.poetic.2016.05.001

Araz, R. O., Serra, X., & Bogdanov, D. (2024). Discogs-VI: A musical version identification dataset based on public editorial metadata. Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR).

Bachmann, M. (2021). Maxbachmann/RapidFuzz: Release 1.8. 0 (Vol. 10) [Computer software].

Daikoku, H., Ding, S., Sanne, U. S., Benetos, E., Wood, A. L., Fujii, S., & Savage, P. E. (2020). Human and automated judgements of musical similarity in a global sample. PsyArXiv Preprint.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint arXiv:1810.04805.

Du, X., Chen, K., Wang, Z., Zhu, B., & Ma, Z. (2022). Bytecover2: Towards Dimensionality Reduction of Latent Embedding for Efficient Cover Song Identification. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 616–620. https://doi.org/10.1109/ICASSP43922.2022.9747630

Du, X., Wang, Z., Liang, X., Liang, H., Zhu, B., & Ma, Z. (2023). Bytecover3: Accurate cover song identification on short queries. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095389

Du, X., Yu, Z., Zhu, B., Chen, X., & Ma, Z. (2021). ByteCover: Cover Song Identification via Multi-Loss Training (arXiv:2010.14022). arXiv. https://doi.org/10.48550/arXiv.2010.14022

Ellis, D. P. W. (2011). The "covers80" cover song data set. http://labrosa.ee.columbia.edu/projects/coversongs/covers80

Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research, 45(3), 239–251. https://doi.org/10.1080/09298215.2016.1200631

Flexer, A., & Lallai, T. (2019). Can we increase inter-and intra-rater agreement in modeling general music similarity? Conference of International Society for Music Information Retrieval (ISMIR), 494–500.

Flexer, A., Lallai, T., & Rašl, K. (2021). On evaluation of inter- and intra-rater agreement in music recommendation. Transactions of the International Society for Music Information Retrieval, 4(1), 182–194. https://doi.org/10.5334/tismir.107

Ghosh, S., Sperling, R., & Hooper, S. (2019). Using Amazon MTurk for research in academia: A beginner's guide for using Qualtrics, detecting VPN/proxy, limiting countries using geolocation & other tips. SSRN. https://doi.org/10.2139/ssrn.3455722

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & others. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv Preprint arXiv:2005.08100.

Hachmeier, S., Jäschke, R., & Saadatdoorabi, H. (2022). Music Version Retrieval from YouTube: How to Formulate Effective Search Queries?. In *LWDA* (pp. 213-226).

Hanson, J. (2018). Assessing the educational value of YouTube videos for beginning instrumental music. Contributions to Music Education, 43, 137–158.

Hu, S., Zhang, B., Lu, J., Jiang, Y., Wang, W., Kong, L., Zhao, W., & Jiang, T. (2022). WideResNet with Joint Representation Learning and Data Augmentation for Cover Song Identification. Interspeech 2022, 4187–4191. https://doi.org/10.21437/Interspeech.2022-10600

Jones, M. C., Downie, J. S., & Ehmann, A. F. (2007). Human similarity judgments: Implications for the design of formal evaluations. International Society for Music Information Retrieval Conference (ISMIR).

Li, B., & Kumar, A. (2019). Query by video: Cross-modal music retrieval. International Society for Music Information Retrieval Conference (ISMIR).

Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W.-C. (2020). Deep entity matching with pre-trained language models. arXiv Preprint arXiv:2004.00584.

Liikkanen, L. A., & Salovaara, A. (2015). Music on YouTube: User engagement with traditional, user-appropriated, and derivative videos. Computers in Human Behavior, 50, 108–124. https://doi.org/10.1016/j.chb.2015.01.067

Liu, F., Tuo, D., Xu, Y., & Han, X. (2023). CoverHunter: Cover song identification with refined attention and alignments. 2023 IEEE International Conference on Multimedia and Expo (ICME), 1080–1085.

Martet, S. (2016). The circulation of user-appropriated music content on YouTube. YouTube and Music, 22(4), 169.

Matherly, T. (2018). A Panel for Lemons? Positivity bias, reputation systems and data quality on MTurk. European Journal of Marketing, 53. https://doi.org/10.1108/EJM-07-2017-0491

McDaniel, B. (2021). Popular music reaction videos: Reactivity, creator labor, and the performance of listening online. New Media & Society, 23(6), 1624–1641. https://doi.org/10.1177/1461444820918549

Mellis, A., & Bickel, W. (2020). Mechanical turk data collection in addiction research: Utility, concerns, and best practices. Addiction (Abingdon, England), 115. https://doi.org/10.1111/add.15032

Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. arXiv Preprint arXiv:1803.10963.

Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behavior Research Methods, 46, 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Silva, D. F., de Souza, V. M., & Batista, G. E. (2015). Music shapelets for fast cover song recognition. ISMIR, 441–447.

Smith, J. B. L., Hamasaki, M., & Goto, M. (2017). Classifying derivative works with search, text, audio, and video features. International Conference on Multimedia and Expo (ICME), 1422–1427. https://doi.org/10.1109/ICME.2017.8019444

Xu, X., Chen, X., & Yang, D. (2018). Key-invariant convolutional neural network toward efficient cover song identification. 2018 IEEE International Conference on Multimedia and Expo (ICME), 1–6. https://doi.org/10.1109/ICME.2018.8486531

Yesiler, F., Doras, G., Bittner, R. M., Tralie, C. J., & Serrà, J. (2021). Audio-based Musical Version Identification: Elements and Challenges. IEEE Signal Processing Magazine, 38(6), 115–136. https://doi.org/10.1109/MSP.2021.3105941

Yesiler, F., Serrà, J., & Gómez, E. (2020a). Accurate and scalable version identification using musically motivated embeddings. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 21–25. https://doi.org/10.1109/ICASSP40776.2020.9053793

Yesiler, F., Serrà, J., & Gómez, E. (2020b). Less is more: Faster and better music version identification with embedding distillation (arXiv:2010.03284). arXiv. https://doi.org/10.48550/arXiv.2010.03284

Yesiler, F., Tralie, C., Correya, A., Silva, D. F., Tovstogan, P., Gómez, E., & Serra, X. (2019). Da-TACOS: a dataset for cover song identification and understanding. Proc. of the 20th Int. Soc. for Music Information Retrieval Conf. (ISMIR), 327–334.

Yu, Z., Xu, X., Chen, X., & Yang, D. (2020). Learning a representation for cover song identification using convolutional neural network. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 541–545. https://doi.org/10.1109/ICASSP40776.2020.9053839

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Exploring the role of large language model in collaborative travel planning task

Hikaru Kumamoto and Hideo Joho

DOI: https://doi.org/10.47989/ir30iConf46966

Abstract

Introduction. This study explores how generative AI, specifically ChatGPT, influences collaborative travel planning. Understanding its effect on tasks like trip planning reveals insights into human-AI collaboration, particularly how AI tools support decision-making and streamline information gathering.

Method. Twenty participants (10 pairs) planned a 1-night, 2-day trip under two conditions: (1) using only Google, and (2) using Google with ChatGPT. This within-subject study measured completion time, satisfaction, and plan quality via questionnaires and observation to capture task performance, user behaviour, and collaboration dynamics.

Analysis. Data was analysed using t-tests and Wilcoxon signed-rank tests to compare completion times, satisfaction, and plan quality. Analyses of conversation volume and ChatGPT logs provided insights into AI-assisted collaboration dynamics and interaction patterns.

Results. No significant difference in task completion time was found. However, plans made with ChatGPT were more complete and aligned with requirements. Participants found information more easily with ChatGPT, but satisfaction levels remained similar, suggesting that easier information access did not translate to higher satisfaction.

Conclusions. Generative AI improves collaborative search task quality but does not enhance efficiency or satisfaction. AI tools like ChatGPT are effective for providing structured information and are best used as complementary resources alongside traditional search engines in planning tasks.

Introduction

Navigating the sheer volume of available data has been both a challenge and an opportunity. The ability to collaborate effectively on information-seeking tasks has become essential, particularly for complex undertakings like travel planning, where diverse sources of information must be integrated to accommodate various preferences and constraints (Shar, 2014). Collaborative decision-making presents unique challenges not encountered in individual tasks, such as time inefficiencies due to group deliberation, coordination complexities, and the influence of dominant group members, which can skew outcomes (Kraemer & King, 1988). Understanding these dynamics is crucial for designing effective tools for collaborative settings. The emergence of generative AI models, especially Large Language Models like OpenAI's ChatGPT, is transforming this landscape. These systems provide information, synthesise ideas, and respond conversationally to users' prompts based on patterns identified in extensive pre-existing human-generated data. By addressing the unique challenges of collaborative tasks, this study investigates how generative AI might enhance complex processes like trip planning, where collective decision-making is central.

Generative AI tools like ChatGPT offer unique potential in collaborative contexts. Unlike search engines, which require users to manually sift through and interpret results, ChatGPT engages interactively, providing suggestions, organizing information, and helping users think through problems. This shift from passive search to active collaboration raises possibilities for improving task efficiency, reducing users' cognitive burden, and enhancing the quality of outcomes (Gienapp, 2024). Despite this potential, the impact of AI on collaborative information-seeking, especially in practical settings like travel planning, remains underexplored (Skjuve, et al., 2023). Travel planning, with its multiple stages—from destination selection to booking accommodations—offers an ideal context to examine how AI might influence collaborative processes and outcomes (Imazu, et al., 2011; Morris, 2013).

To explore these dynamics, this study poses the following key research questions:

- How does the use of ChatGPT, impact the efficiency, user satisfaction, and quality of travel plans in collaborative information-seeking tasks compared to traditional search engine use?
- What influence does ChatGPT have on communication dynamics and decision-making during collaborative tasks, and what challenges arise when integrating AI into real-time, dynamic information needs like travel planning?

To address these questions, the study examines the role of ChatGPT in real-world collaborative travel planning tasks. Specifically, it investigates the effect of incorporating ChatGPT into the planning process, comparing it with the traditional use of search engines. While prior research has explored AI's effectiveness in individual information-seeking tasks, little attention has been given to its role in collaborative efforts, where multiple users must coordinate to reach decisions. Travel planning, an everyday activity, reflects the complexity of many collaborative tasks, involving negotiation, compromise, and logistical decision-making. Understanding how AI supports such processes has broader implications for the future of collaboration across various fields.

Methodology

The methodology of this research involved a within-subjects experimental design, allowing participants to complete a travel planning task under two conditions: (1) using only a search engine (condition S), and (2) using both a search engine and ChatGPT (condition C). The purpose was to evaluate task efficiency, satisfaction, and the quality of travel plans produced. The order of conditions was counterbalanced to control for learning effects.

We used a free edition of ChatGPT version 3.5 (OpenAI, 2022), the latest available during the study in August 2023.

Participants

Ten pairs of participants (20 individuals) were recruited for this study. Participants were recruited in pairs and chose partners they felt comfortable collaborating with. This self-selection ensured they did not have difficulty working together, minimizing potential influence of interpersonal dynamics on the results. The participants consisted of undergraduate and postgraduate students from the University of Tsukuba, with ages ranging from 19 to 27 years (mean age = 21.3 years). The sample included 6 women and 14 men, with academic backgrounds spanning informatics, science and engineering, social sciences, humanities, and medicine. Participants were daily users of traditional search engines, ensuring basic familiarity with online search methods such as finding relevant information and comparing sources. Advanced search techniques (e.g., Boolean operators or relevance ranking) were not specifically assessed, as the study aimed to reflect typical user behaviour. Before the experiment, participants were briefed on ChatGPT's interface, capabilities, and general use cases to standardise their understanding of the tool.

Task

The task asked participants to collaboratively plan a two-day, one-night domestic trip. The planning process was divided into four key steps:

- Destination selection: participants chose three tourist destinations they would like to visit.
- Schedule creation: participants organised the trip itinerary, including the sequence of visits to destinations.
- Accommodation booking: participants selected and booked a place to stay.
- Transport booking: participants arranged transportation for the trip, including modes of travel and approximate timings.

Participants were asked to start with the destination selection but could order the remaining steps as they wished. They could not return to a previous procedure, as this would make time recording unreliable.

In condition S (search engine only), participants used Google to gather information and complete the task without AI assistance. In condition C (search engine + ChatGPT), participants used both ChatGPT and Google, with the frequency of ChatGPT use left to their discretion. However, the study did not record the precise duration of ChatGPT use in place of the search engine. Future studies could measure this to better understand time allocation across tools and its impact on performance. For consistency, participants were not allowed to use other AI-based tools or services.

Time	Place	Activity	Transportation	Duration	Price
7:00	City A to City B		Local Train	80 min	840
...					
11:10	Market C	Shopping, lunch		60 min	
...					
17:30	Hotel D	Check-in, break		60 min	3776
...					

Table 1. Template of travel planning task

Procedure
Each experiment session lasted approximately 170 minutes, divided into two phases (condition S and condition C). Participants were given 80 minutes per condition to complete the four steps of travel planning. After each condition, they completed a post-task questionnaire assessing satisfaction with the process and the travel plan produced.

During the experiment, one participant in each pair operated the computer, while the other recorded the travel plan (See Table 1). Participants could switch roles as needed. Importantly, all decisions regarding the travel plan had to be made collaboratively, ensuring the study focused on group dynamics rather than individual decision-making.

Data collection
Data were collected on three key variables:

- Task completion time: the total time participants took to complete the travel planning task was recorded for each condition. Additionally, the time spent on each of the four steps was tracked.
- Satisfaction: after each condition, participants completed a questionnaire using a 5-point Likert scale, rating satisfaction with the planning process, ease of use, and the quality of the travel plan. Questions addressed both overall satisfaction and satisfaction with specific stages of the planning process.
- Travel plan quality: the travel plans were evaluated based on criteria such as completeness (whether the plan covered all essential aspects of the trip), feasibility (whether the schedule and budget were realistic), and attention to detail (whether transportation, accommodation, and destination information were sufficiently detailed).

Additionally, the volume of conversation during the task was recorded and analysed to assess ChatGPT's impact on communication. Conversations were transcribed and measured in terms of the number of words exchanged, both between participants and with ChatGPT.

Ethical considerations
This study was approved by the Ethics Review Committee of Institute of library, information and media science, University of Tsukuba (Approval No. 23-73). All participants provided informed consent before participating in the experiment and were compensated with Amazon gift certificates based on the institution's regulation.

Results
The results of the study are presented in three key areas: task completion time, user satisfaction, and the quality of the travel plans produced. Additionally, we analysed the volume of conversation between participants and ChatGPT's usage in condition C (search engine + ChatGPT). Each of these aspects is compared between the two conditions—condition S (search engine only) and condition C.

Task completion time
Overall task completion time includes the time taken to plan the entire two-day trip. The median task completion time for condition C was 3,553 seconds (59:13 minutes), while condition S had a median time of 3,471 seconds (57:51 minutes). The Wilcoxon rank-sum test indicated that this difference was not statistically significant ($p = 0.50$), suggesting that including ChatGPT did not reduce overall task completion time.

However, when broken down by task procedure, a few notable differences emerged. In the destination selection stage (procedure A), participants in condition C took slightly less time than

those in condition S. The median time for procedure A in condition C was 21.0% of the total time, while in condition S it accounted for 29.9%. This indicates that ChatGPT may have helped participants generate ideas for destinations more quickly. Conversely, the schedule creation stage (procedure B) took slightly longer in condition C (33.1%) compared to condition S (26.0%), though these differences were not statistically significant ($p > 0.05$).

User satisfaction

Participants' satisfaction with the travel planning process was measured using a post-task questionnaire, which included questions on ease of use, satisfaction with the plan produced, and satisfaction with each stage of the planning process. Overall, there were no significant differences in total satisfaction between the two conditions. Participants generally expressed moderate satisfaction in both conditions, with a median satisfaction score of 4 (on a 5-point Likert scale).

However, analysis of individual stages revealed nuanced differences. In the destination selection stage, participants in condition C reported finding it easier to identify relevant information compared to condition S (median satisfaction score of 1.00 vs. 2.00). In contrast, in the accommodation booking stage (procedure C), participants in condition S expressed greater satisfaction with their ability to find accurate and up-to-date information. This difference likely reflects ChatGPT's reliance on static data, limiting its effectiveness for tasks requiring real-time information, such as finding available accommodations.

Despite the overall similarity in satisfaction between the two conditions, open-ended responses from participants highlighted ChatGPT's usefulness for idea generation. One participant remarked, '*ChatGPT helped us brainstorm destinations quickly when we had no initial ideas*,' while another stated, '*the search engine required more effort to manually gather all the information.*'

Quality of travel plans

The quality of the travel plans produced was assessed based on completeness, feasibility, and level of detail. Plans were considered complete if they covered all essential aspects of the trip, including transportation, accommodation, and activities. Feasibility referred to whether the plans were realistic in terms of time and budget, and level of detail was evaluated based on how thoroughly the plans addressed logistical aspects, such as transport schedules and booking details (Imazu, et al., 2011).

Results showed that condition C (search engine + ChatGPT) led to travel plans that were more aligned with the task requirements. Nine of the ten pairs in condition C produced plans that met all task criteria (e.g., stayed within budget, covered all trip details), compared to seven pairs in condition S. Additionally, plans from condition C were often rated as more detailed, especially in terms of destination choices and activities. However, condition S produced more feasible schedules, as participants found it easier to obtain accurate transport and accommodation information using search engines.

For example, one travel plan in condition C proposed a tight schedule including three distant locations in a single day, which may have been unrealistic. In contrast, travel plans in condition S generally reflected more conservative itineraries, with better alignment between travel times and planned activities.

Conversation volume and ChatGPT usage

The volume of conversation between participants was measured to assess how ChatGPT influenced group interaction. In condition C, participants exchanged fewer words compared to condition S, suggesting that ChatGPT may have reduced the need for extended discussions (Joho, et al., 2008). The median number of words exchanged in condition C was approximately 15% lower than in condition S. A Wilcoxon rank-sum test found no statistically significant difference between

the conditions ($p = 0.34$), but the trend suggests the AI tool streamlined some aspects of the decision-making process.

Participants used ChatGPT primarily in the destination selection (procedure A) and schedule creation (Procedure B) stages. The most common type of interaction with ChatGPT was asking for ideas and itinerary suggestions. For instance, participants frequently asked ChatGPT for recommended places to visit in a specific region or to help build a draft schedule. ChatGPT was used less frequently in accommodation and transport booking stages, likely due to the need for real-time data, which ChatGPT could not provide, at the time of study.

Conclusive discussion

This study investigated the impact of generative AI, specifically ChatGPT, on collaborative travel planning tasks by comparing it to traditional search engine use. Collaborative decision-making differs from individual decision-making due to challenges such as reconciling diverse preferences, negotiating compromises, and managing group dynamics, including the influence of dominant members (Kraemer & King, 1988; Shah, 2014). These attributes make collaborative tasks particularly complex and critical for evaluating the potential of AI tools. While no statistically significant differences were found in overall task completion time or satisfaction, the results provide nuanced insights into how AI tools influence collaborative information-seeking behaviour and decision-making. These findings highlight both the potential and limitations of AI integration into real-world tasks, offering several important takeaways for future research and practical applications.

Key findings

The most notable benefit of using ChatGPT in the collaborative planning process was in the destination selection stage, where it helped participants generate ideas quickly. In this phase, participants using ChatGPT took less time to decide on destinations compared to those relying solely on search engines. This suggests that generative AI can serve as an effective tool for brainstorming, particularly in tasks requiring creativity or decision-making when users lack initial direction. Participants noted that ChatGPT helped overcome the initial inertia of decision-making by offering suggestions and guiding the conversation. Participant satisfaction ratings did not indicate concerns about ChatGPT's suggestions being overly leading. This aligns with research suggesting that AI's value lies in its ability to generate ideas and reduce cognitive load in early problem-solving stages (Maaravi, et al., 2021).

However, ChatGPT's advantages did not extend uniformly across all planning stages. In the accommodation and transport booking stages, participants in condition S (search engine only) expressed greater satisfaction and produced more feasible plans. This can be attributed to ChatGPT's lack of real-time data, which limited its ability to provide accurate, up-to-date information on hotel availability, transportation schedules, and other logistics-dependent tasks. In these cases, search engines offered more reliable information, underscoring the continued importance of traditional tools for tasks heavily reliant on real-time data. ChatGPT's structured, system-driven outputs may also not align with the adaptability and creativity humans bring to dynamic contexts.

Despite these insights, several limitations must be acknowledged. First, the sample size was relatively small, with only 10 pairs of participants. This limited the statistical power of the results, potentially explaining the lack of significant differences in task completion times and satisfaction across conditions. A larger sample size could provide more robust data and reveal subtler trends in performance and behaviour.

Additionally, participants were primarily students with limited experience using ChatGPT for travel planning. This familiarity bias may have limited their ability to fully leverage the tool's capabilities, particularly in the task's initial stages. Future studies should include a longer

familiarisation period or recruit participants with greater experience using AI tools, allowing a more accurate assessment of how generative AI can be used effectively in collaborative settings.

Another limitation was the task's time constraint, which many participants found restrictive. The one-hour limit may have prevented participants from fully exploring travel planning options, especially in condition C, where ChatGPT unfamiliarity slowed some processes. Extending task duration in future studies would provide a more realistic reflection of how AI can be integrated into collaborative work, particularly for more complex or fluid tasks like travel planning.

Implications for the iSchool community

For researchers and practitioners in the iSchool community, this study provides valuable insights into the evolving role of AI in information-seeking behaviours. AI tools like ChatGPT can complement, but not replace, traditional search engines in collaborative tasks. Combining AI for idea generation with search engines for real-time data retrieval is a promising approach, particularly in complex tasks like travel planning that require both creativity and logistical accuracy.

The findings also suggest that AI's impact on group communication and dynamics warrants further investigation. While AI may reduce cognitive load and streamline decision-making, it risks diminishing the collaborative aspect of tasks, potentially leading to less robust decision-making. Future work should explore how to balance AI assistance with human interaction to preserve the benefits of collaboration while leveraging AI's strengths.

Conclusion

In conclusion, this study found that while ChatGPT offers distinct advantages in collaborative travel planning, particularly in the ideation phase, it does not enhance all aspects of the task universally. Search engines remain critical for tasks requiring real-time data and detailed logistics. Additionally, while AI may reduce cognitive load and streamline decision-making processes, it changes the nature of collaboration by reducing the need for extended discussions. These findings suggest that generative AI tools are best used as complementary resources, handling early-stage ideation and exploration, while traditional tools support detailed decision-making.

Future research should address the study's limitations by expanding the sample size and including participants with more diverse backgrounds and experience levels to improve the findings' generalisability. Investigating the use of generative AI across a broader range of collaborative tasks—such as project management or academic research—would offer further insights into its applications. To address limitations related to real-time data, future studies could explore integrating AI with live databases or APIs to enhance decision-making accuracy. Lastly, extending task durations in future experiments could allow participants to better engage with AI tools, providing a deeper understanding of leveraging these technologies for long-term, complex collaborative tasks.

Acknowledgements

This work was supported in part by JSPS KAKENHI grant number JP24K03048, Japan. The authors thank the study participants for their valuable contributions and the anonymous reviewers for their constructive feedback, which greatly improved this work.

About the authors

Hikaru Kumamoto was a fourth-year undergraduate student at the College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba, at the time of this work. She received the dean of the School of Informatics award in recognition of her outstanding academic performance.

Hideo Joho is a Professor of Information Retrieval in the Institute of Library, Information and Media Science, University of Tsukuba. His research focuses on understanding human information interactions, building search user interfaces, and developing test collections for evaluation. He can be reached at hideo@slis.tsukuba.ac.jp.

References

Shah, C. (2014). Collaborative information seeking. Journal of the Association for Information Science and Technology, 65(2), 215-236.

Kraemer, K.L. & King, J.L. (1988). Computer-based systems for cooperative work and group decision making. ACM Computing Surveys, 20 (2), 115–146.

OpenAI (2022). Introducing ChatGPT. https://openai.com/index/chatgpt/

Morris, M. R. (2013). Collaborative search revisited. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1181-1192).

Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023). The user experience of ChatGPT: findings from a questionnaire study of early users. In *Proceedings of the 5th international conference on conversational user interfaces* (pp. 1-10).

Gienapp, L., Scells, H., Deckers, N., Bevendorff, J., Wang, S., Kiesel, J., ... & Potthast, M. (2024). Evaluating generative ad hoc information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1916-1929).

Imazu, M., Nakayama, S., & Joho, H. (2011). Effect of explicit roles on collaborative search in travel planning task. In *Information Retrieval Technology: 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings 7* (pp. 205-214). Springer Berlin Heidelberg.

Joho, H., Hannah, D., & Jose, J. M. (2008). Comparing collaborative and independent search in a recall-oriented task. In *Proceedings of the second international symposium on Information interaction in context* (pp. 89-96).

Maaravi, Y., Heller, B., Shoham, Y., Mohar, S., & Deutsch, B. (2021). Ideation in the digital age: literature review and integrative model for electronic brainstorming. *Review of Managerial Science*, 15, 1431-1464.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Information places: myth, affect, and gaps in information retrieval modelling

Nathan R. Johnson

DOI: https://doi.org/10.47989/ir30iConf46960

Abstract

Introduction. This article contributes to theory building for information theory through a critical-historical approach to the ancient Simonides story, a foundational myth of human memory. By examining the Simonides myth, which details orality-based memory practices, the study identifies underemphasized aspects of memory in the social transitions to literacy.

Method. A critical-historical approach is used to examine the contemporary relevance of this influential memory myth in relation to information theory research and emerging technologies.

Analysis. The analysis highlights how users' affective and contextual experiences provide valuable information for designing information systems that deliver more precise and relevant results.

Results. The article argues that modeling the context of a user's affective experience enhances the effectiveness of information systems. It introduces and refines key concepts, including informational economy, judgement, symbol-affect, and informational alibi.

Conclusions. The author suggests further refinement and application of the concepts of informational economy, judgement, symbol-affect, and informational alibi to improve the design of information retrieval systems, emphasizing their potential to align systems more closely with embodied experience.

Introduction

Modern foundational theories in Information Retrieval (IR), predominantly based on three theoretical models—exact match, vector space, and probabilistic—tend to parallel the conduit model of information, which has been critiqued for its oversimplified view of information as merely transmitted rather than created or interpreted (Day, 2000). While these models provide robust frameworks for IR systems, they can overlook crucial aspects of information behaviour, notably the emotional, sensory, and tacit dimensions that significantly influence user interaction and information reception (Cox et al., 2017; Marcella et al., 2019; Nahl & Bilal, 2007). These issues have been recognized, but few alternative models have been advanced (Frohmann, 1990; Hjørland & Nissen Pedersen, 2005; Jansen & Rieh, 2010). Recognizing and naming these less studies elements of IR is essential for advancing theory and enhancing the practical toolkit of professionals in the field.

To enrich current theoretical approaches and operationalize these overlooked dimensions, this research revisits an ancient narrative: the myth of Simonides. This myth, influential in shaping historical and modern conceptions of memory and information behaviour (Yates, 1966), recorded the method of loci, a mnemonic technique often interpreted as an early model of known item-information retrieval. The method of loci involves cognitively associating information with specific physical locations within a familiar environment, such as a house or a path, allowing users to connect these spaces mentally to retrieve stored information. Despite its antiquity, the interpretation of this myth has had significant theoretical implications for both psychological models of memory and the modelling of IR systems (Yates, 1966).

Scholars have noted that the historical interpretation of the Simonides myth has frequently been condensed into overly simplistic frameworks. By reinterpreting this story, this research illuminates and defines key concepts critical for IR research, such as relevance and precision—concepts that presuppose users will recognize retrieval results as relevant they resemble previously known ideas. These assumptions underpin many IR techniques but do not encompass the full spectrum of challenges in retrieving information. Although numerous experts in IR have acknowledged these limitations, the conceptual toolkit for IR theory remains underdeveloped (Dobreski et al., 2021), at least in terms of pedagogy and practice. By reinterpreting the Simonides myth, this study highlights concepts that are pivotal for understanding human memory and, consequently can be usefully deployed for designing more effective information retrieval systems with similar models. Although speculative, this analysis encourages ongoing development of IR theory that integrates these richer, more nuanced dimensions of information behaviour. More specifically, it introduces four conceptual ideas that refine existing IR scholarship: economy, judgement, symbol-affect, and informational alibi.

Method

To make its argument, this study uses a conceptual approach to reinterpret the myth of Simonides in relation to modern information retrieval systems, specifically focusing on the integration of affect and community relationships. By doing so, it draws from approaches to information systems that understand users from communal perspectives rather than psychological ones (Star, 1989). This research scrutinizes classical texts and their commentaries to uncover how the ancient mnemonic practices can inform current IR technologies. It's critical reading is primarily driven by affect theory, which encourages a hermeneutical approach that focuses on the movement of bodies and sensations instead of overemphasizing text and information (Ahmed, 2004; Ahmed et al., 2009). This reading is similar to previous hermeneutical approaches in information studies that have used critical-historical approaches to revisit taken-for-granted concepts (Buckland, 1991, 1997; Day, 2000, 2011, 2014; Frohmann, 1992, 2004, 2004; Madsen, 2016). The theoretical analysis produces critical review of secondary sources that discuss Simonides' method of loci, supplemented by contemporary theories of affect and information theory. This method allows for

a nuanced understanding of the limitations and capabilities of IR in replicating the affective components of human memory and information processing. The study juxtaposes a newer interpretation of the ancient mnemonic strategies with modern IR applications to illustrate the gap between human affective memory and IR's theory.

The research data for this study consisted of primary and secondary sources related to the Simonides myth of memory, a foundational narrative in the history of information retrieval (IR). These sources highlighted the weaknesses of historical assumptions that memory operates as a stable and retrievable repository, a concept that has guided the development of IR modelling. By critically reading those texts, I traced an alternative view of memory and IR. My analysis was guided by current critiques in IR research, which highlight significant gaps in the field, particularly concerning the dynamic and contextual nature of human memory and its representation in digital systems. This combined approach enabled me to connect the myth's enduring legacy to ongoing challenges and opportunities for innovation in IR theory and practice.

In the next section, this paper describes the basic Simonides myth and points out several of its weaknesses. In the following section, I interpret secondary texts that contribute new perspectives of the Simonides myth, specifically those that identify contextual issues that would have informed early interpretations of the Simonides myth. In the last section, I offer four theoretical concepts that the myth highlights while suggesting that operationalizing those terms can help further IR research.

Analysis

Simonides of Ceos, a canonical figure among the nine canonical Greek lyric poets, lived between 556 and 468 BCE, positioning him as a precursor to classical philosophers like Socrates and Aristotle (Rawles, 2018a). He relocated from Ceos to Attica around 526 BCE, drawn by the financial prospects for his poetry. His career trajectory shifted following the assassination of his initial patron, Hipparchus, leading him to Thessaly where he gained the support of notable families like the Scopadae (Rawles, 2018b). Simonides' method of loci is exemplified in his well-documented banquet incident at the court of Scopas. Commissioned to compose an ode in honour of a boxer, Simonides included references to the Dioscuri—twin gods revered in athletic circles—which Scopas criticized as unsuitable. Subsequently, Scopas withheld payment, suggesting that Simonides seek his reward from the twins themselves. This event took a dramatic turn when, during the banquet, Simonides was called outside by two youths. In his brief absence, the banquet hall collapsed, killing all inside but leaving Simonides unharmed. The poet's subsequent role in identifying the deceased by recalling their seating positions at the table led to the development of the method of loci. The context of why Simonides was at the banquet is important for understanding his method of loci.

Ancient treatises on memory, like those by Cicero and Quintilian, often focus only on the latter part of the Simonides myth, specifically the method of loci (Cicero, 1954; Quintilian, 1922). In their rhetoric handbooks, they used this technique to distinguish between natural and artificial memory—a concept that parallels has historically informed modern information theory and subsequent IR systems, particularly those concerned with human computer interaction (Hayles, 1999). Natural memory refers to our innate cognitive abilities, while artificial memory relies on structured methods, like the method of loci, to enhance recall. This classical distinction highlights the limitations of AI in replicating the deeply contextual and emotionally nuanced processes involved in human memory and information processing. The Simonides method remains relevant today, offering insights into the complex relationship between information, user, and system—elements that remain challenging for IR systems to fully emulate. Even in modern IR, the boundary between artificial and natural memory continues to shape system design and functionality as systems are conceptualized as separate from users rather than as part of the same ecosystem (Latour, 1990; Mongili & Pellegrino, 2014). If a "natural" user is interfacing with an "artificial" system,

it is difficult to imagine IR as anything other than an exercise in precision and relevance, because the "artificial" system necessarily contains a '*correct*' answer that a user can ideally identify. Much of the contemporary research in information studies drawing from actor-network theory approaches are advances in seeing IR as enmeshed with users (Tatnall, 2005).

Theoretical approaches similar to the method of loci has had a significant impact on 20th-century information theories, which underpin much of modern information studies, including IR (Day, 2000). Conceptually, information is difficult to pin down, largely because it's imagined as a container for other phenomenon (Day, 2000). In practice, this has led to most studies imagining information as '*thing*' (Buckland, 1991). Information as thing highlights mutable objects like referring to data, documents, and other physical objects that can convey knowledge. One of the reasons this thing framework is popular for developing theory is that other definitions are largely ambiguous, at least to date (Furner, 2014).

The Simonides myth and method of loci offer important perspectives for IR. Greek and Roman authors expanded upon the method of loci, introducing two critical principles that are often overlooked. First, they argued that specific words should not be associated with memory places, a theoretical technique that is intellectually similar to Salton's vector space model that depends on specific entities and their relationships (Salton et al., 1975). Cicero and Quintilian both stressed the importance of remembering ideas or vivid images over exact words, largely because they believed as many preliterate cultures did, that words gestured toward general feelings (Rubin, 1995). Although Cicero acknowledged that the method could be used for word-for-word recall, he recommended it only for practice (Cicero, 1942, p. 2.87.357). The second principle was that memory places must be set against striking, memorable backdrops. For example, the *Rhetorica ad Herenium* advises that exceptional events like a lunar eclipse should serve as memory cues, as they evoke a stronger emotional response (Cicero, 1954, p. 3.22). Together, these principles suggest that the art of memory relies on enhancing natural memory through vivid imagery, placed in an easily recalled, unique background. Over time, these ideas have been adapted to various memory practices, from notetaking to public memory studies and cognitive theories of information behaviour.

However, contemporary interpretations of the Simonides myth have identified important limitations of the original interpretations (Rawles, 2018b). For instance, Simonides associated banquet seats with individual identities, implying that ordinary locations could serve as loci. This contradicts Cicero and Quintilian's advice to use striking, visually rich images as memory aids. Furthermore, Simonides did not order the loci himself; the banquet seating was predetermined. This contradicts the myth's emphasis on arranging loci in a deliberate order. The story also blurs the line between natural and artificial memory, as Simonides stumbled upon the method by accident—his recollection wasn't a trained skill, but a consequence of the roof's collapse. The usual interpretation of the myth lends itself to supporting precision and relevance theories, but those could also be supplemented with theoretical concepts that further develop artificial/natural as a mutually dependent concept where artificial and natural aren't in opposition but are considered asymmetries in systems design. Moreover, the myth would highlight how important vividness is for information retrieval. One might imagine how much easier it is to search for a topic or object that is more emotionally forceful.

The Simonides myth also points to several alternative perspectives for memory due to details that have been left underexamined. These include the following questions: Why is the detail of Scopas short-changing Simonides included if the story is meant to illustrate a simple mnemonic technique? And why are Castor and Pollux invoked, despite their irrelevance to mnemonic methods? Additionally, both Cicero and Quintilian express doubts about the practical effectiveness of the method of loci, as Quintilian sceptically noted (Yates, 1966, p. 24). This raised the question that numerous historians later addressed: Cicero, Quintilian, and the model of the method of loci had been interpreted too simplistically. New scholarship has provided more nuanced readings of

Cicero and Quintilian (Leff, 1996). Cicero and Quintilian both noted that Simonides was not just the inventor of memory palaces, he was also renowned as a sophist poet with a reputation for being preoccupied with the economics of poetry. Some of the method of loci story was to provide a story about memory practices that points to Simonides as a much more elaborate theorist of information and memory. Much of his work emphasized the importance of compensation for artistic labour (Rawles, 2018b, p. 15) and his obsession with payment was so well known that he became a target for ridicule (Rawles, 2018b, pp. 138–139). Indeed, one of Simonides' contemporaries, the Greek poet Aristophanes, even used him as a symbol of greed in *Clouds*, comparing Sophocles' desire for wealth to Simonides' legendary avarice (Aristophanes, 1998, p. 690). Similarly, Aristotle's *Rhetoric* presents a dialogue where Simonides asserts that wealth is more valuable than wisdom, as the wise serve the wealthy (Aristotle, 1926, p. 1391a). This economic concern is tacit in the memory palace retellings by Cicero and Quintilian and is critical for understanding the relationship between Simonides and his patron, Scopus. Simonides, having never been paid what he was owed, is saved by divine intervention, suggesting that the gods compensated him instead.

The inclusion of Castor and Pollux in the myth offers another clue to the economic themes underlying the story. Quintilian notes that the involvement of the Dioscuri is likely fictional, yet their presence is significant. Both Cicero and Quintilian introduce Castor and Pollux when the topic of payment arises: first when Scopus refuses to pay Simonides, and again when the twins save the poet, symbolically '*paying their debt*'. Beyond narrative allusion, there is historical evidence linking the Dioscuri to money. The first Roman denarius, minted centuries after Simonides, depicted Castor and Pollux on horseback (Meadows & Williams, 2001, p. 38). While Simonides would not have known this Roman coinage, the association of the twins with money could have developed over time, transforming the memory and economic themes in popular retellings of the story.

A deeper analysis suggests that Simonides' callers—allegedly the Dioscuri—were not mythical gods but metaphorical symbols for something more worldly important for the method of loci. Neither Cicero nor Quintilian provides an alternate explanation for why Simonides left the banquet hall just before the roof collapsed. The coincidence of Simonides' timely departure is usually interpreted as divine intervention or the result of shoddy architecture. However, if Castor and Pollux represent money, they may point to a different cause for the massacre. Simonides, an unpaid poet and the sole survivor, may have taken revenge on Scopus, the client who refused to pay him. The method of loci becomes part of Simonides' alibi, providing a reason for him to be in Scopus' home (contracted to deliver a panegyric) and a convenient excuse for his survival (he was called away). The mnemonic device, in this reading, is a ruse—Simonides fabricates it to identify the dead and secure contracts for their eulogies, using memory to serve his own economic interests.

This interpretation diverges from the traditional reading of the method of loci as an art of memory. Quintilian doubts the practical utility of the technique, implying that it may be little more than a parlour trick. Simonides never had to prove the accuracy of his mnemonic device; he could have confused the identities of the dead, and no one would have known. Quintilian's scepticism hints at a richer, more sophisticated art of memory, one that transcends the mnemonic technique used in Simonides' alibi. The inclusion of the Dioscuri in this story highlight potential the importance of the economy in the production of memory and information. Simonides' preoccupation with payment for his poetry points to a concept of memory as transactional or economic. In IR, this could be likened to the '*cost*' of retrieval—whether in terms of computing resources, relevance, or user attention—making efficiency and precision crucial elements of the process.

Building on the reinterpretation of Simonides' method of loci, it is insightful to contrast his broader mnemonic practices with those of his contemporary, the lyric poet Sappho. Unlike Simonides, whose approach to memory was structured and precise, Sappho's method was deeply embedded in evoking powerful emotional responses. Her memory practices were not concerned with

accuracy but with creating an atmosphere that cultivated feelings of serenity and beauty through pastoral landscapes and religious spaces. This focus on affect and atmosphere provides an alternative perspective on the method of loci, emphasizing the emotional and ethical dimensions of memory rather than mechanical recall (Jarratt, 2002, p. 20). Such an approach underscores the importance of the loci's emotional resonance—a factor that cognitive science affirms as crucial in memory retention (McGaugh, 2013; Reisberg & Hertel, 2004). Sappho's flexible and fluid use of memory loci introduces a third concept for information retrieval (IR): the ethical positioning of memory. This concept explores how the selection and application of memory (or information) in IR can shape outcomes and influence moral judgments. For IR scholarship, this raises important ethical considerations about the retrieval process, the use of information, and its broader societal impacts. Additionally, Sappho's perspective recontextualizes Quintilian's rendition of the Simonides myth. Quintilian acknowledges the familiarity of the story, suggesting that readers fill in the gaps (Quintilian, 1922, p. 11.2.11). He recounts Simonides' alleged motivation to secure payment from his patron, possibly through coercive means. This interpretation notes that Simonides may have orchestrated the collapse of Scopus' dining hall, using his mnemonic method as an alibi to avenge unpaid debts and ensure future employment. In this light, the method of loci becomes more than a tool for memory—it transforms into a strategic device for achieving justice and retribution.

Quintilian's description enhances this interpretation by noting the mnemonic power of physical places to evoke past interactions, emotions, and unresolved conflicts. When Simonides revisits the location of his dispute with Scopus, not only does he recognize the place, but he also vividly recalls the slight and his ensuing moral outrage. Memory, then, serves as a '*symbol of navigation*', guiding Simonides through a landscape charged with ethical implications and leading him to recall not just events, but also the emotional weight they carried. This navigational aspect of memory uses symbols—like the Dioscuri representing money—to evoke the loci of betrayal and poetic justice.

The ethical narrative extends into Cicero's interpretation, where memory is described not merely as a technique for recalling facts but as a method for imposing order and moral reckoning on past injustices. Cicero suggests that the effective use of memory involves selecting meaningful loci that align with the facts one wishes to remember, thus preserving both the sequence of events and their ethical (Cicero, 1942, p. 2.86.354). Contrary to a simplistic understanding of Simonides' memory technique, Cicero portrays it as a complex interplay of memory, ethics, and oratory, where the loci selected are imbued with the moral values and philosophical insights gained through a lifetime of experience and education.

A closer reading and contextualization of Simonides along with Cicero's and Quintilian's description of the poet highlight how memory emerges from an affective orientation shaped by education and ethical considerations. The art of memory is not just about linking names to seats but about using one's memory to navigate the philosophical and ethical challenges of life, emphasizing the role of oratory in cultivating a public voice that resonates with personal convictions and societal values. This reframing lends itself to several models that can enrich theory development in IR. The following section uses the hermeneutical methods from this section to propose concepts to supplement current IR research.

Results

This section describes theoretical concepts that can be implied from the Simonides myth and how they might inform new studies in IR. The interpretation in this paper lends itself to four concepts that can be useful for developing theory in IR. These are offered as ways to think about how to measure and operationalize important parts of IR systems, both those that exist and those be developed. Many of these concepts have existing research that can further contribute to developing them as a part of IR modelling.

Economy

The concept of '*economy*' refers to the transactional nature of memory, particularly how memories are formed, stored, and recalled as a form of currency in both cognitive and social interactions. Most IR research that takes up the idea of information economy largely focuses on information as economic commodity in a larger market of goods (Porot, 1998; Varian, 1995). In the narrative involving Simonides, the Mnemonic Economy is vividly illustrated through his preoccupation with the economics of poetry and his legendary concern for the compensation of his artistic work. His method of loci isn't just a tool for remembering but a strategy to negotiate and secure payment. The memory technique serves as a medium through which the value of memory and knowledge is negotiated—both literally, in terms of Simonides securing his dues, and metaphorically, to assert the worth of his intellectual labour. For example, '*economy*' can be seen in how advertising works as part of information behaviour, where brands create memorable slogans or jingles that consumers recall at the point of purchase. Simonides' myth highlights how memory produces the conditions of information value. Information is influenced both by economic interests while also encouraging new economic interests, as a strong, memorable brand identity can drive consumer behaviour and increase sales.

Judgment

'*Judgment*' in IR would refer to the ability to discern and recall information that is pertinent to one's current context or needs, reflecting a sophisticated form of cognitive judgment. Simonides' use of the method of loci exemplifies this concept as he strategically remembers and organizes information based on its relevance to his situational and emotional context. When revisiting the place of his previous humiliation, Simonides recalls not only the physical space but also the emotional landscape of his interaction, including the unresolved conflict and his moral outrage. This selective recall guided by relevance to his goals (securing justice and reparation) underscores the inherent judgment involved in what is remembered and what is left aside. A juror in a trial uses judgment over information when they focus on specific testimony or evidence that aligns with the legal questions they must consider in their verdict. This selective memory encourages them to ignore some information and focus on what is most critical for their decision-making process. Judgment as a concept may be particularly important as new IR systems that integrate generative AI transform where and when acts of judgment occur (Hersh, 2024).

Symbol-affect

The concept of '*symbol-affect*' highlights how information is not only about recalling discrete facts but also about invoking the emotional and symbolic meanings attached to those memories. In your narrative, the places and symbols (like the Dioscuri as representations of money and betrayal) are imbued with deep emotional significance that informs Simonides' recall process. These loci are not just spatial or visual markers but are charged with affective energy that colours the information and influences the recall process. Simonides' emotional response to these symbols guides his recollection and actions, demonstrating how information can be a deeply affective experience, shaped by the emotional resonance of the symbols and places encoded in one's memory. A war veteran might have a strong emotional reaction when visiting a war memorial, which serves as a symbolic loci triggering memories not just of the events, but of the emotions felt during those times—fear, camaraderie, loss. This example shows how locations steeped in personal or collective emotion can dramatically shape memory retrieval, specifically regarding user relevance (Saracevic, 2017).

Informational alibi

Informational alibi involves the use of mnemonic techniques not just for recall but as a strategic defence or justification for one's actions, embedding a deeper narrative purpose within the mnemonic practice. In the story, Simonides crafts an alibi through his method of loci, using it to justify his presence at and departure from Scopus' banquet. This alibi serves a dual purpose: it

provides a cover story that exonerates him from involvement in the catastrophic events and strategically positions him to claim a moral high ground in the aftermath. The informational alibi thus becomes a tool for narrative construction, where information serves not only as a recall mechanism but to weave a narrative that aligns with the individual's broader objectives and ethical stance. In legal settings, a witness might reconstruct their memory of an event in a way that aligns with a narrative beneficial to their side of the story. For instance, remembering and emphasizing details that support their innocence, or the guilt of others can serve as a mnemonic alibi, strategically shaping the narrative recounted in court. The concept of informational alibi can help illuminate IR concepts related to system transparency (Bernard & Balog, 2023).

Conclusion

The above critical reading of the Simonides myth introduces new concepts important for advancing IR. This research's analysis extends beyond the traditional understanding of the method of loci as a simple mnemonic device, proposing a more complex interaction between memory, ethics, and economics that transcends the purely cognitive functions typically associated with IR systems. The story of Simonides, as reinterpreted here, does not merely illustrate a method for enhancing memory recall but also exposes the ethical and economic dimensions that can influence memory practices. These dimensions are not merely side effects of IR systems, they are critical for understanding how IR and users work together. Simonides' strategic use of memory for securing justice and recognition, along with the symbolic representation of money through the Dioscuri, underscores memory's role not only as a cognitive process but also as a tool for navigating social and moral landscapes. This deeper exploration into the motivations behind memory usage introduces new theoretical concepts for IR, such as mnemonic economy, relevance and judgment, symbolic loci and affect, and mnemonic alibi, enriching the standard IR paradigms of precision and relevance.

By recontextualizing the method of loci through the ethical and affective lenses provided by Sappho and Quintilian, the study highlighted the complex interplay between memory and value systems. This approach suggests that IR systems could benefit from integrating these broader dimensions to more effectively mimic human cognitive and emotional processes. The historical depth provided by Simonides' experiences offers profound insights into the affective dimensions of memory that modern IR technologies still struggle to replicate or understand. In retelling this narrative, this paper does not merely revisit a historical method but revitalizes it, proposing that the principles underlying ancient mnemonic techniques can inform and expand contemporary IR research. This discussion opens new avenues for considering how IR systems might incorporate not only the technical mechanisms of memory but also the ethical and emotional considerations that profoundly shape human memory practices.

The Simonides myth, including its issues dealing with information, ethics, and economics, offers valuable lessons for enhancing the sophistication of IR systems and deepening our understanding of the role of memory in society. This exploration encourages a continuous dialogue on the limitations and potentials of IR, urging a re-evaluation of how technology interprets and engages with human memory and knowledge. As this research navigated the intersection of ancient wisdom and modern technology, the enduring relevance of classical memory practices and their value to IR is increasingly apparent, guiding future innovations in IR toward more sophisticated analysis.

Acknowledgements

The author thanks the two anonymous reviewers who provided critical feedback that significantly improved the manuscript.

About the author

Nathan R. Johnson is an Associate Professor at the University of South Florida in Tampa, Florida USA. He received his Ph.D. from the University of Wisconsin—Madison's Information School. His research interests are in the history of information, specifically on how infrastructure and labour shape technological systems. He can be contacted at nathanjohnson@usf.edu.

References

Ahmed, S. (2004). Collective feelings: Or, the impressions left by others. *Theory, Culture & Society*, 21(2), 25–42. https://doi.org/10.1177/0263276404042133

Ahmed, S., Massumi, B., Probyn, E., & Berlant, L. (2009). *The affect theory reader* (M. Gregg & G. J. Seigworth, Eds.). Duke University Press.

Aristophanes. (1998). *Peace* (S. D. Olson, Trans.). Clarendon Press.

Aristotle. (1926). *Art of rhetoric* (J. H. Freese, Trans.). Harvard University Press.

Bernard, N., & Balog, K. (2023). A systematic review of fairness, accountability, transparency and ethics in information retrieval. *ACM Computing Surveys*. https://doi.org/10.1145/3637211

Blair, C., Dickinson, G., & Ott, B. L. (2010). Introduction: Rhetoric/memory/place. In G. Dickinson, C. Blair, & B. L. Ott (Eds.), *Places of Public Memory: The Rhetoric of Museums and Memorials* (pp. 1–54). University of Alabama Press.

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science*, 42(5), 351–360. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

Buckland, M. K. (1997). What is a "document"? *Journal of the American Society for Information Science*, 48(9), 804–809. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<804::AID-ASI5>3.0.CO;2-V

Cicero. (1954). *Rhetorica ad Herennium* (H. Caplan, Trans.). Harvard University Press.

Cicero, M. T. (with Rackham, H.). (1942). *De oratore* (E. W. Sutton, Trans.). Harvard University Press.

Cox, A. M., Griffin, B., & Hartel, J. (2017). What everybody knows: Embodied information in serious leisure. *Journal of Documentation*, 73(3), 386–406. https://doi.org/10.1108/JD-06-2016-0073

Day, R. E. (2000). The "conduit metaphor" and the nature and politics of information studies. *Journal of the American Society for Information Science*, 51(9), 805–811. https://doi.org/10.1002/(SICI)1097-4571(2000)51:9<805::AID-ASI30>3.0.CO;2-C

Day, R. E. (2011). Death of the user: Reconceptualizing subjects, objects, and their relations. *Journal of the American Society for Information Science and Technology*, 62(1), 78–88. https://doi.org/10.1002/asi.21422

Day, R. E. (2014). *Indexing it all: The subject in the age of documentation, information, and data*. MIT Press.

Dobreski, B., Zhu, X., Yang, T., & Ridenour, L. (2021). Information organization and information retrieval in the LIS curriculum. *Proceedings of the Association for Information Science and Technology*, 58(1), 707–709. https://doi.org/10.1002/pra2.536

Frohmann, B. (1990). Rules of indexing: A critique of mentalism in information retrieval theory. *Journal of Documentation*, 46(2), 81–101. https://doi.org/10.1108/eb026855

Frohmann, B. (1992). The power of images: A discourse analysis of the cognitive viewpoint. *Journal of Documentation*, 48(4), 365–386. https://doi.org/10.1108/eb026904

Frohmann, B. (2004). *Deflating information: From science studies to documentation*. University of Toronto Press.

Furner, J. (2014). Information without information studies. In F. Ibekwe-SanJuan & T. M. Dousa (Eds.), *Theories of Information, Communication and Knowledge* (Vol. 34, pp. 143–179). Springer Netherlands. https://doi.org/10.1007/978-94-007-6973-1_7

Hayles, K. (1999). *How we became posthuman: virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Hersh, W. (2024). Search still matters: Information retrieval in the era of generative AI. *Journal of the American Medical Informatics Association*, 31(9), 2159–2161. https://doi.org/10.1093/jamia/ocae014

Hjørland, B., & Nissen Pedersen, K. (2005). A substantive theory of classification for information retrieval. *Journal of Documentation*, 61(5), 582–597. https://doi.org/10.1108/00220410510625804

Jansen, B. J., & Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology*, 61(8), 1517–1534. https://doi.org/10.1002/asi.21358

Jarratt, S. C. (2002). Sappho's memory. *Rhetoric Society Quarterly*, 32(1), 11–43. https://doi.org/10.1080/02773940209391219

Latour, B. (1990). Technology is society made durable. *The Sociological Review*, 38(S1), 103–131. https://doi.org/10.1111/j.1467-954X.1990.tb03350.x

Leff, M. (1996). Commonplaces and argumentation in Cicero and Quintilian. *Argumentation*, 10(4), 445–452. https://doi.org/10.1007/BF00142977

Madsen, D. (2016). Liberating interdisciplinarity from myth. An exploration of the discursive construction of identities in information studies. *Journal of the Association for Information Science and Technology*, 67(11), 2697–2709. https://doi.org/10.1002/asi.23622

Marcella, R., Baxter, G., & Walicka, A. (2019). User engagement with political "facts" in the context of the fake news phenomenon. *Journal of Documentation*, 75(5), 1082–1099. https://doi.org/10.1108/JD-11-2018-0180

McGaugh, J. L. (2013). Making lasting memories: Remembering the significant. *Proceedings of the National Academy of Sciences*, 110(Supplement 2), 10402–10407. https://doi.org/10.1073/pnas.1301209110

Meadows, A., & Williams, J. (2001). Moneta and the Monuments: Coinage and Politics in Republican Rome. *The Journal of Roman Studies*, 91, 27–49. https://doi.org/10.2307/3184768

Mongili, A., & Pellegrino, G. (2014). *Information infrastructure(s): Boundaries, ecologies, multiplicity*. Cambridge.

Nahl, D., & Bilal, D. (2007). *Information and emotion: The emergent affective paradigm in information behavior Research and Theory*. Information Today.

Porat, M. U. (1998). The information economy: Definition and measurement. In J. Cortada (Ed.), *Rise of the Knowledge Worker* (pp. 101–131). Butterworth-Heinemann.

Quintilian. (1922). *Institutio oratoria* (H. E. Butler, Trans.). Harvard University Press.

Rawles, R. (2018a). Introduction. In *Simonides the poet: Intertextuality and reception* (pp. 1–20). Cambridge University Press.

Rawles, R. (2018b). *Simonides the poet: Intertextuality and reception*. Cambridge University Press.

Reisberg, D., & Hertel, P. (Eds.). (2004). *Memory and emotion*. Oxford University Press.

Rubin, D. C. (1995). *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*. Oxford University Press.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. https://doi.org/10.1145/361219.361220

Saracevic, T. (2017). *The notion of relevance in information science: Everybody knows what relevance is. But, what is it really?* Morgan & Claypool.

Star, S. L. (1989). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In L. Gasser & M. N. Huhns (Eds.), *Distributed Artificial Intelligence, Volume II* (pp. 37–54). Pitman. https://dx.doi.org/110.1016/B978-1-55860-092-8.50006-X

Tatnall, A. (2005). Actor-Network Theory in Information Systems Research. In *Encyclopedia of Information Science and Technology, First Edition* (pp. 42–46). IGI Global. https://doi.org/10.4018/978-1-59140-553-5.ch009

Varian, H. R. (1995). The information economy. *Scientific American*, 273(3), 200–202.

Yates, F. (1966). *The art of memory*. University of Chicago Press.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Detecting text-level intellectual influence with Knowledge graph embeddings

Lucian Li and Eryclis Silva

DOI: https://doi.org/10.47989/ir30iConf46957

Abstract

Introduction. Tracing the spread of ideas and the presence of influence is a question of special importance across a wide range of disciplines, ranging from intellectual history to cultural analytics, computational social science, and the science of science.

Method. We collect a corpus of open-source journal articles, generate Knowledge Graph representations using the Gemini LLM, and attempt to predict the existence of citations between sampled pairs of articles using previously published methods and a novel graph neural network based embedding model.

Results. We demonstrate that our knowledge graph embedding method is superior at distinguishing pairs of articles with and without citation. Once trained, it runs efficiently and can be fine-tuned on specific corpora to suit individual researcher needs.

Conclusions. This experiment demonstrates that the relationships encoded in a knowledge graph, especially the types of concepts brought together by specific relations can encode information capable of revealing intellectual influence. This suggests that further work in analysing document level knowledge graphs to understand latent structures could provide valuable insights.

Introduction

Tracing the spread of ideas and the presence of influence is a question of special importance across a wide range of disciplines, ranging from intellectual history to cultural analytics, computational social science, and the science of science. While some fields (such as bibliometrics) generally rely on annotated citation datasets, such data is limited to past few decades and the domain of scholarly publication. A wide range of potential questions rely on solely on unstructured text.

Previous attempts to explore and discover examples of influence in corpora have not been previously validated on annotated corpora, so we have no sense of the false negative rate. We propose a novel method to calculate pairwise influence between documents and evaluate it against previously published methods.

We generate individual knowledge graphs for each document and then calculate Knowledge Graph embeddings. We experiment with node level embeddings like TransE (Bordes et al, 2013) but find that the computational costs of generating embeddings using the combined knowledge graphs across our article corpus proved prohibitive. Other approaches, such as anchoring, proved infeasible because of the relatively small size of individual knowledge graphs.

Our final approach utilizes Graph Convolutional Networks to train an embedding model through the contrastive task of identifying pairs of knowledge graphs generated using the Gemini LLM. The intuition is that each knowledge graph represents a conceptual space emerging from arguments and claims advanced in the document and that structural and semantic similarities in knowledge graphs correspond to latent structural similarities in the text and argumentative style.

Related work

Influence detection

With the large-scale digitization of archives, the past decade has seen a range of different approaches to detect influence and similarity across documents. These have relied on three families of methods broadly: text reuse, topic models, and word sense similarity.

Text reuse methods, such as (Funk and Mullen, 2018) and (Cordell et al., 2024), essentially use n-gram overlap to detect large runs of text shared between documents. While able to detect all instances of direct quotation, these methods cannot detect paraphrase or indirect influence.

Next, several studies, such as (Rockmore et al., 2018) and (Barron et al., 2018) have compared topic model distributions across documents. While these methods capture high level changes in subjects and disciplinary focus, they are unable to evaluate specific changes in argumentation or offer any insight into intersections in semantic content.

Finally, another family of recent studies focus on detecting similarities in word sense, either by calculating similarities in word embeddings for the same word across different documents as in (Soni et al., (21) or calculating the BERT perplexity for unexpected words as in (Vicinanza et al., 2023). These methods focus on influence in the domain of language use and word choice. They can capture stylistic influence and innovative language, but again fail to capture semantic or argumentative similarity.

Lastly, some studies have leveraged knowledge graphs to calculate conceptual similarity, although these methods use knowledge graphs to encode information rather than calculating similarity in knowledge graph structure (Zhang and Zhu, 2022) (Bi et al., 2022).

Knowledge graph generation

Previous triplet extraction tools include traditional NLP based entity and relation extraction pipelines such as Kim et al. (2021) and LLM based approaches like Melnyk et al. (2021), Bronzini et

al. (2023), Sun et al. (2024), and Han et al. (2023). We experimented with locally hosted solutions such as Rebel and Mistral 7b but found subjectively that the triplets and resulting knowledge graphs produced by Google's Gemini were more representative and convincing. We eventually used Google's Gemini Team et al. (2023) LLM because of its relatively high accuracy, free access, and relatively low implementation complexity.

Knowledge graph embeddings

A range of projects, such as Li et al. (2023), have worked on using knowledge graph embeddings to calculate similarity. These, however, generally focus on creating large single knowledge graphs from documents, citations, and metadata. Our method, which aims to align multiple independent knowledge graphs, is closer to approaches such as Fanourakis et al. (2023), which compares embeddings across different knowledge graphs to align entities. Although their goal is to combine different KG, this shows that embeddings across independent KG may be comparable. Baumgartner et al. (2023) and Huang et al. (2022) also attempt to align entities across different KG, but they find that base embeddings may need additional transformations or anchoring to be effectively compared across different embedding spaces. Our final approach also draws from previous attempts to create a Graph Convolutional Network based embedding system that encodes graph structure through contrastive learning, applied in knowledge graphs by Xu et al. (2019). We apply a generally similar training architecture, but incorporate a different strategy to select contrastive pairs and incorporate additional semantic data in node features

Knowledge graph evaluation metrics

Knowledge graph (KG) evaluation is a critical area of research that has seen various approaches proposed in recent years. These approaches can be broadly categorized into structure-based, data quality-based, and task-oriented evaluation methods.

Structure-based evaluation metrics have been a primary focus of several studies. Seo et al. (2022) presented novel metrics such as the Instantiated Class Ratio, Instantiated Property Ratio, Class Instantiation, Subclass Property Acquisition, and Subclass Property Instantiation. They also proposed the Inverse Multiple Inheritance metric to assess ontology complexity. These metrics provide valuable insights into the structural quality of knowledge graphs.

Complementing structure-based approaches, Dou et al. (2023) proposed different measures of knowledge in KGs, introducing more broadly applicable metrics. Their K Score, I Score, and C Score, derived from the science of science, information theory, and causality perspectives respectively, offer a multifaceted approach to KG evaluation. This work aims to address limitations in existing structure-based and data quality-based assessment techniques.

Moving beyond intrinsic evaluation methods, Heist et al. (2023) presented a framework to evaluate KGs via downstream tasks. This framework enables a comprehensive assessment of KGs by evaluating their performance on multiple kinds of tasks such as classification, regression, and recommendation. Their experiments revealed significant variations in KG performance depending on the specific task, highlighting the importance of extrinsic evaluation metrics as a complement to established intrinsic criteria.

These diverse approaches demonstrate the complexity of KG evaluation and the need for multifaceted assessment methods that consider both intrinsic quality and practical utility in downstream applications.

Method

A flowchart overview of the methods is provided in Figure 1.

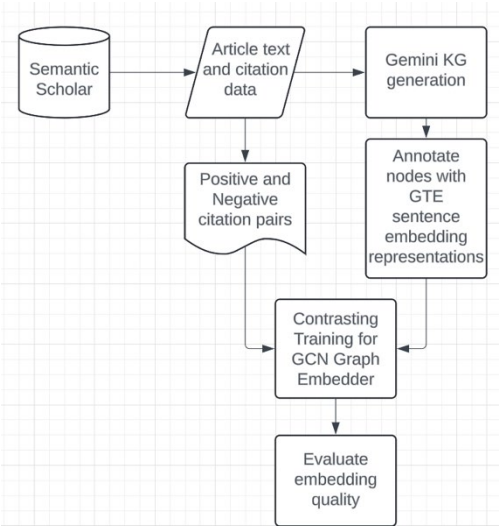


Figure 1. Methods flowchart

Dataset

As an evaluation dataset, we used the Semantic Scholar API to download pairs of recent academic articles on specific subjects (Darwinism, terrorism, convolutional graph learning, carbon offset, and MECP2, a neurological disease gene). These subjects were selected to cover a wide range of disciplines to evaluate the ability of the method to generalize across disciplines instead of fitting to norms of one area.

We downloaded the text of roughly 4,400 open-source articles divided evenly into about 900 per subject.

From these articles we sampled pairs. We generate roughly 22,000 total pairs, of which 8,500 are '*positive samples*' where one article cites the other. We also generate 13,500 negative samples, which are pairs of articles about the same subject that do not cite each other. This was done to prevent the model from learning only to distinguish article topics. Imbalance in the negative and positive pairs was deliberately chosen as both types of relationships are able to inform the model. This is generally in line with previous contrastive training objectives.

Knowledge graph generation

We initially explored using the Mistral 7B model locally for our knowledge graph generation task. However, through subjective evaluation, we found that Gemini Pro consistently produced more meaningful and relevant entity-relationship triples compared to Mistral 7B.

Given the dataset at hand, our approach with Gemini Pro entailed systematically sampling one-third of the dataset to assess the model's efficacy in extracting entities and relationships. We began by partitioning the textual data, carefully considering the context window constraints to optimize model performance. Utilizing LangChain we employed recursive character segmentation, fine-tuning parameters such as size and overlap.

After meticulously customizing prompts for each text chunk, we applied the Gemini Pro model. Challenges arose during entity extraction due to the presence of numerous citations, which prompted us to refine the prompt, instructing the model to disregard direct citations. As a result, we obtained JSON lists containing two nodes and one edge each, which were then utilized to enhance the original data frame.

In our approach with Gemini Pro, we adopted a one-shot learning paradigm to train our model, a machine learning technique where a model learns to recognize patterns or make predictions based

on only a single example or a few examples of each class. By leveraging one-shot learning, our model can effectively extract entities and relationships from the dataset with minimal labelled examples, enhancing efficiency and reducing the need for extensive data annotation. We provide a visualization of one document's encoding in a knowledge graph in Figure 2.

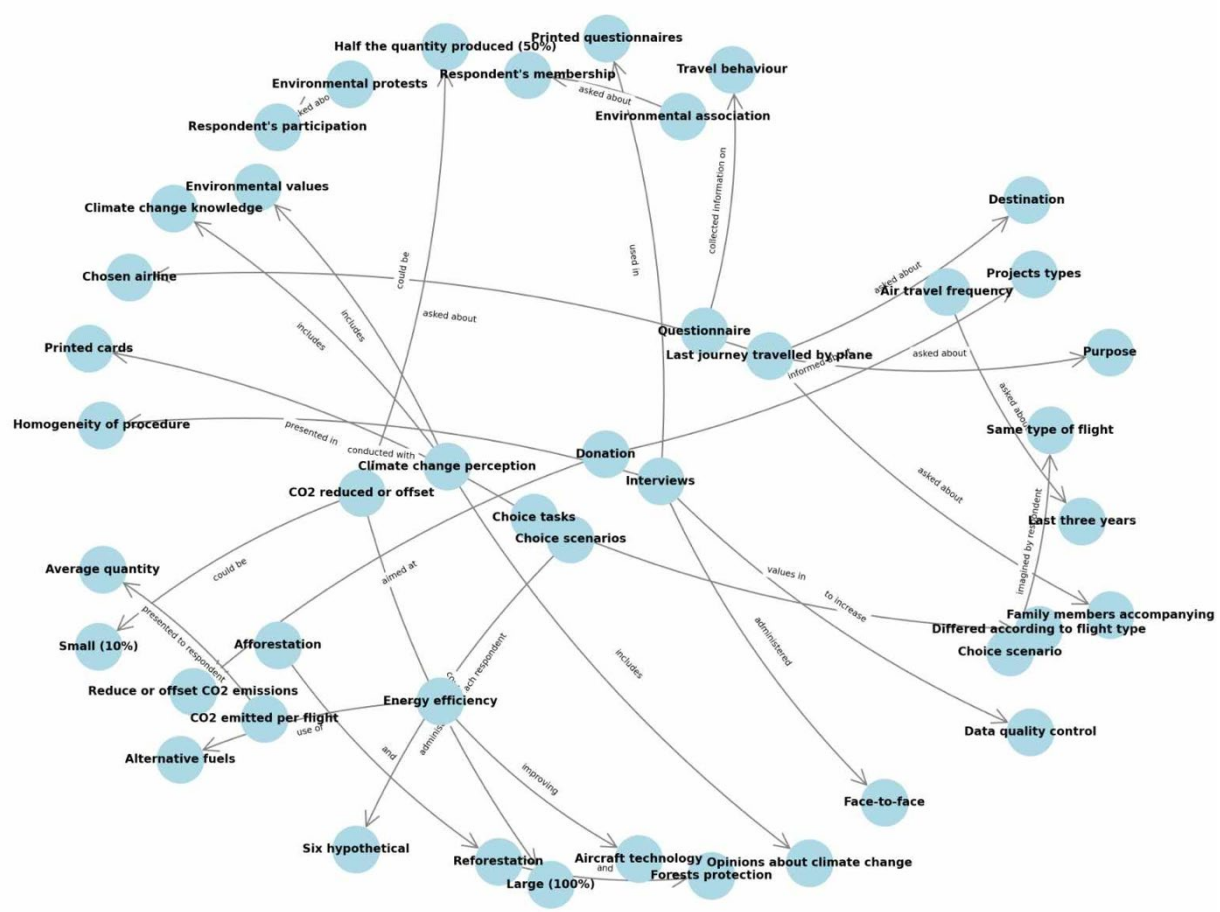


Figure 2. Sample knowledge graph generated from Gemini

Analysis

Graph neural network embedding

Using our positive and negative sampled pairs, we conducted contrastive training using a three-layer graph convolutional network (GCN).

For the node features, we generate semantic embeddings using GTE Li et al. [2023], a BERT based sentence embedding method. These 384-dimension semantic representations of the node content are used as node features. In theory, including these representations as node attributes will allow the model to treat nodes conveying similar concepts in the same way.

We initialize 500 dimensions for the hidden layer, and 100 dimensions for the final graph embedding. the 100-dimensional final layer representations were mean pooled into one 100-dimensional graph embedding. Loss was calculated with the pyTorch CosineEmbeddingLoss function, which penalizes high distance between true pairs and low distance between false pairs. Loss was backpropagated and optimized with the Adam optimizer.

Currently, hyperparameter tuning through grid search is infeasible because of computational demands, although for future work, we hope to further evaluate other combinations of

hyperparameters and GNN architectures. We use the DGL implementation of GCN with a pyTorch backend.

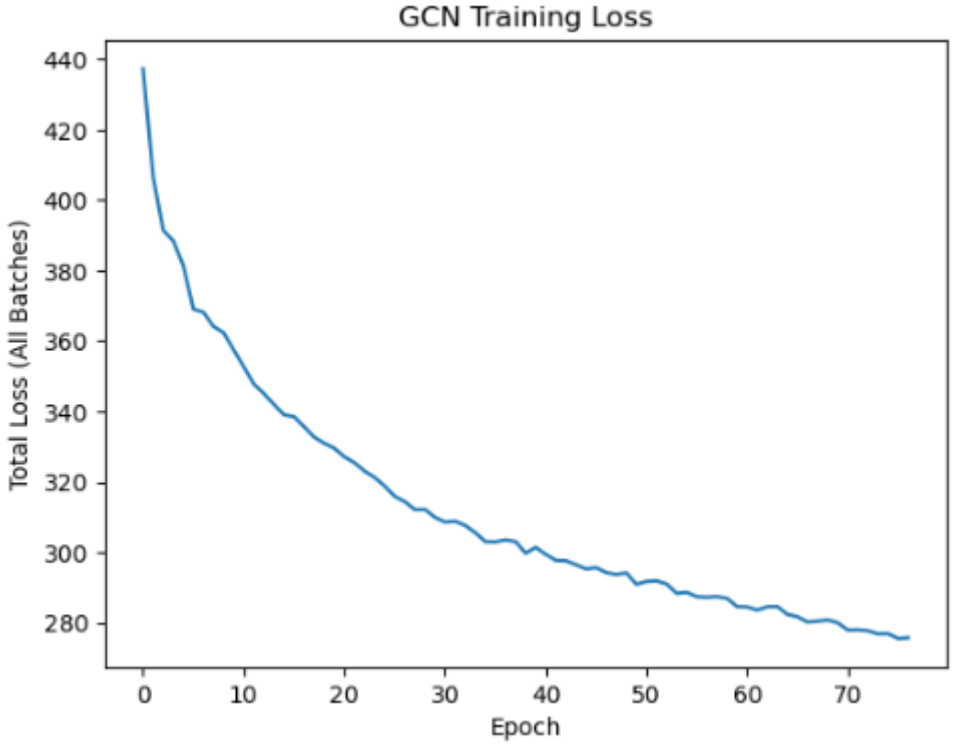


Figure 3. Training error in GCN training

In Figure 3, we observe near convergence after 80 epochs of training. Other hyperparameters are contrastive margin=0.5, learning rate=0.05 and an 80% train-test split.

We observe near convergence after 80 epochs of training. Other hyperparameters are contrastive margin=0.5, learning rate=0.05 and an 80% train-test split.

We also experimented with a Graph Attention Network to weight different types of relationships differently, but the computational demands proved too high for our limited resources.

The final model is able to take a knowledge graph as input and output a 100-dimensional embedding encoding argumentative structure.

Reproduction of previously published approaches

We implement three other methods for a performance comparison. For text reuse detection, we use the text-matcher package Reeve [2020]. We use default settings and remove the first 300 and last 2000 characters to remove titles and listed citations to limit the comparison to the argument text. Text reuse pairs were scored by raw number of overlapping segments. The topic model was implemented with Gensim. After some hyperparameter tuning, we used LDA with 500 topics as that produced the best results. LDA vectors were scored using KL divergence. Finally, GTE embeddings were implemented with the HuggingFace sentence-transformers package. We created 1000 characters chunks, which were embedded and then summed to generate a document embedding. GTE vector pairs were evaluated using cosine distance.

Results

Evaluation metrics:

Because there is no gold standard way to evaluate influence, we used the presence of citation as a proxy metric. The overall goal for evaluation is the citation prediction task. Given two articles about the same topic, we will evaluate the ability of models to predict if a citation exists between the two articles. We report three performance metrics: we compare the distribution of scores reported by each method for true and false pairs. We use a rank-sum test to evaluate the likelihood that the scores for true pairs are significantly higher than those for false pairs. We also calculate the Receiver Operating Characteristic curve for each method. This shows potential trade-offs in performance when each method is used as a classifier to distinguish true and false pairs. Finally, we calculate the optimal threshold for each method using the ROC curve and evaluate the F1-score performance of each method. These metrics capture the distribution of scores produced as well as their potential as a classification model.

Performance

We observe superior performance for our model on all metrics. While all previously published models had significant ability to discriminate true pairs than false pairs, our model achieved high significance of separation for the rank sum metric. This suggests that the separation between the scores for true and false pairs is superior using the KG embeddings.

Metric	Text Reuse	Topic Model KL Divergence	Document Embedding (GTE) distance	KG Embedding (Ours)
Rank sum test (inverse log p-value)	75	191	267	**611**
Area under ROC curve	0.55	0.58	0.59	**0.61**
F1 score (optimal threshold)	0.17	0.46	0.48	**0.55**

Table 1. Accuracy metric comparison

Table 1 shows comparison of our method's accuracy with a range of previously published approaches. Figure 4 shows that our model has a higher area under the ROC, and furthermore dominates the previous models at all points of the curve. At the optimal threshold, our method achieves a higher F1 accuracy score than any published method. Most prior methods achieve sub 0.5 F1-score, suggesting that this is inherently a very difficult problem.

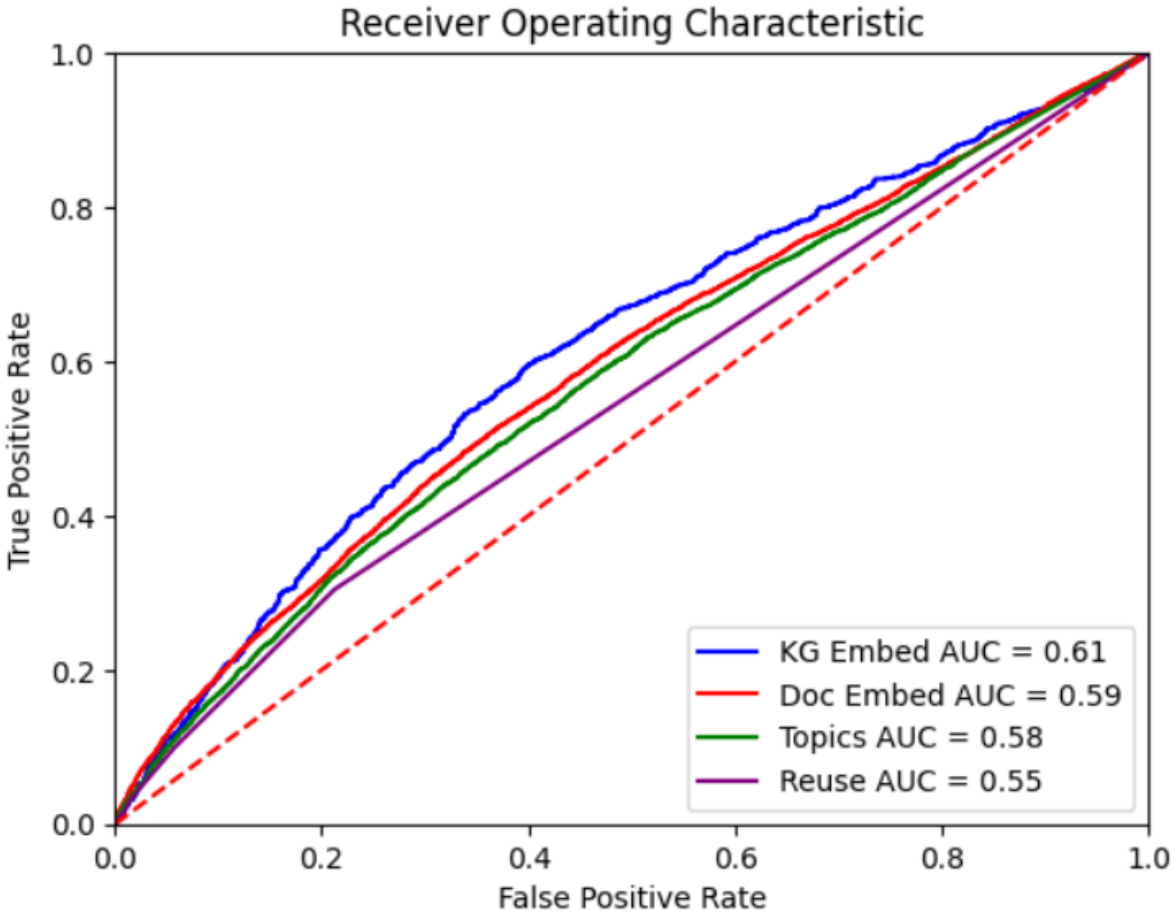


Figure 4. ROC curves for different methods. Our method in blue is superior across all accuracy thresholds.

As demonstrated in Figure 5, the text reuse method specifically performed worse because of the very high number of both true and false documents with no exact text reuse present at all. Embedding based methods, like sentence embeddings and KG embeddings can capture a wider range of similarity with greater granularity.

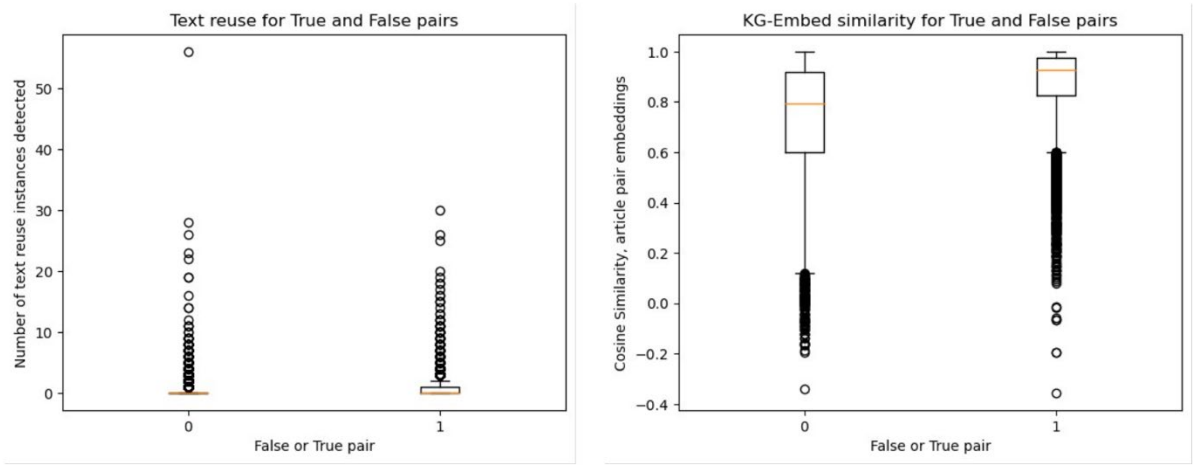


Figure 5. Text reuse vs embedding similarity across documents

Future directions

Currently, the most direct next step is to experiment with locally hosted LLMs to extract KG triplets, especially since great improvements have been made since the original experiments. Especially with the potential to fine tune with human generated triplet examples, we hope to achieve higher quality KG representations. We also hope to streamline the embedding pipeline, especially to experiment with new hyperparameters.

In terms of analysis, we hope to expand this method to domain specific questions of influence in future publications. As reviewers have noted, we will apply this to a more comprehensive cross section of academic publications, especially on historical texts, where real relationships between documents are not clearly documented. We are also interested in exploring other contexts with structured citations, particularly in a legal context with patents and law codes.

Conclusion

We report a knowledge graph-based method for detecting influence that outperforms previously published methods. Our method is trained through a contrastive approach to distinguish citation pairs from non-citation pairs of similar articles. This approach incorporates graph/argumentative structure and the semantic information encoded in knowledge graphs.

One major issue we faced was lack of access to GPU acceleration and compute credits. More complex architectures like Graph Attention Networks, more neural network layers, additional epochs of training, or additional training data would have produced a more effective model. Additional hyperparameter tuning could have produced significantly improved results as well. For future research, additional close reading of successful and unsuccessful cases is necessary. While our method outperforms on quantitative metrics, it is important to understand if it is detecting types of influence previously missed. Potentially, an ensemble method of our KG embedding approach, topic models, and sentence embeddings can achieve even better results by covering for weaknesses for specific models and capturing a wider range of evidence of influence. We also plan to evaluate the method on out of domain data, especially to establish the potential transfer learning ability of the model to generalize to patterns in different contexts.

We have demonstrated that the relationships between nodes and edges in knowledge graphs encode useful information that a neural network can learn to detect latent similarities between documents. We can think of a knowledge graph as representing the conceptual space of a document-our project shows that studying the structures and relationships of this space can provide valuable insights about relationships between documents.

Acknowledgements

We would like to thank Professor Jiaqi Ma for providing feedback and advice on methods and evaluation on this project. We would also like to thank the anonymous reviewers for their valuable comments on future directions and methodological suggestions.

About the authors

Lucian Li is a PhD student at Information School of the University of Illinois. His work focuses on influence detection and similarity detection in unstructured historical/cultural data. He can be contacted at zilul2@illinois.edu

Eryclis Silva is a PhD student at Information School of the University of Illinois. His research focuses on the intersection of law and machine learning, especially fairness in ML models, Legal NLP, and AI Governance. He can be contacted at ersilva2@illinois.edu

References

Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. Individuals, institutions, and innovation in the debates of the French revolution. Proceedings of the National Academy of Sciences, 115(18):4607–4612, 2018.

Matthias Baumgartner, Daniele Dell'Aglio, Heiko Paulheim, and Abraham Bernstein. Towards the web of embeddings: Integrating multiple knowledge graph embedding spaces with fedcoder. Journal of Web Semantics, 75:100741, 2023.

Sheng Bi, Zafar Ali, Meng Wang, Tianxing Wu, and Guilin Qi. Learning heterogeneous graph embedding for Chinese legal document similarity. Knowledge-Based Systems, 250:109046, 2022. ISSN 09507051. doi:https://doi.org/10.1016/j.knosys.2022.109046. URL https://www.sciencedirect.com/science/article/pii/S0950705122005123.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modelling multi-relational data. Advances in neural information processing systems, 26, 2013.

Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. Glitter or gold? deriving structured insights from sustainability reports via large language models. arXiv preprint arXiv:2310.05628, 2023.

Ryan Cordell, David A Smith, Abby Mullen, and Jonathan D Fitzgerald. Going the Rounds Virality in Nineteenth Century American Newspapers. University of Minnesota Press, Forthcoming, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Dou, Jason Xiaotian, Haiyi Mao, Runxue Bao, Paul Pu Liang, Xiaoqing Tan, Shiyi Zhang, Minxue Jia, Pengfei Zhou, and Zhi-Hong Mao. "The Measurement of Knowledge in Knowledge Graphs." In Proceedings of the AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R2HCAI). Washington, DC, USA: Association for the Advancement of Artificial Intelligence (AAAI), 2023.

Nikolaos Fanourakis, Vasilis Efthymiou, Dimitris Kotzinos, and Vassilis Christophides. Knowledge graph embedding methods for entity alignment: experimental review. Data Mining and Knowledge Discovery, 37(5):2070–2137, 2023.

Kellen Funk and Lincoln A Mullen. The spine of American law: Digital text analysis and us legal practice. The American Historical Review, 123(1):132–164, 2018.

Jiuzhou Han, Nigel Collier, Wray Buntine, and Ehsan Shareghi. Pive: Prompting with iterative verification improving graph-based generative capability of llms. arXiv preprint arXiv:2305.12392, 2023.

Nicolas Heist, Sven Hertling, and Heiko Paulheim. 2023. KGrEaT: A Framework to Evaluate Knowledge Graphs via Downstream Tasks. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23). Association for Computing Machinery, New York, NY, USA, 3938–3942. https://doi.org/10.1145/3583780.3615241.

Hongren Huang, Chen Li, Xutan Peng, Lifang He, Shu Guo, Hao Peng, Lihong Wang, and Jianxin Li. Cross-knowledge graph entity alignment via relation prediction. Knowledge-Based Systems, 240:107813, 2022. ISSN 09507051. doi: https://doi.org/10.1016/j.knosys.2021.107813

Vlad-Iulian Ilie, Ciprian-Octavian Truic˘a, Elena-Simona Apostol, and Adrian Paschke. Context-aware misinformation detection: A benchmark of deep learning architectures using word embeddings. IEEE Access, 9:162122–162146,2 021. doi:10.1109/ACCESS.2021.3132502.

Taejin Kim, Yeoil Yun, and Namgyu Kim. Deep learning-based knowledge graph generation for covid-19. Sustainability, 13(4):2276, 2021.

Guangtong Li, L Siddharth, and Jianxi Luo. Embedding knowledge graph of patent metadata to measure knowledge proximity. Journal of the Association for Information Science and Technology, 74(4):476–490, 2023.

Igor Melnyk, Pierre Dognin, and Payel Das. Grapher: Multi-stage knowledge graph construction using pretrained language models. In NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.

Jonathan Reeve. Text-matcher. https://github.com/JonathanReeve/text-matcher, 2020.

Daniel N Rockmore, Chen Fang, Nicholas J Foti, Tom Ginsburg, and David C Krakauer. The cultural evolution of national constitutions. Journal of the Association for Information Science and Technology, 69(3):483–494, 2018.

Seo, Sumin, Heeseon Cheon, Hyunho Kim and Dongseok Hyun. Structural Quality Metrics to Evaluate Knowledge Graphs. ArXiv preprint arXiv:2211.10011, 2022.

Sandeep Soni, Lauren F Klein, and Jacob Eisenstein. Abolitionist networks: Modeling language change in nineteenth century activist newspapers. Journal of Cultural Analytics, 6(1), 2021.

Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. ArXiv preprint arXiv:2312.11805, 2023.

Paul Vicinanza, Amir Goldberg, and Sameer B Srivastava. A deep-learning model of prescient ideas demonstrates that they emerge from the periphery. PNAS nexus, 2(1): pgac275, 2023.

Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. Cross-lingual knowledge graph alignment via graph matching neural network. arXiv preprint arXiv:1905.11605, 2019.

Jinzhu Zhang and Lipeng Zhu. Citation recommendation using semantic representation of cited papers' relations and content. Expert Systems with Applications, 187:115826, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.115826

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Information access via voice commands on YouTube: empirical evidence on the consequences for a marginalised community in Bangladesh

Joy Bhowmik, Viviane Frings-Hessami, Gillian Oliver, and Md. Khalid Hossain

DOI: https://doi.org/10.47989/ir30iConf46945

Abstract

Introduction. This paper examines the socioeconomic and cultural consequences of using voice commands on YouTube to access information by a small-scale marine fishing community in Bangladesh.

Method. The study employs a qualitative approach, grounded in ethnographic fieldwork, to explore the challenges in accessing and utilising information on YouTube. Data were collected through focus groups and interviews with male and female individuals in the community, offering insights into their unique experiences.

Analysis. The data was analysed using reflexive thematic analysis, beginning with categorising the types of information accessed and their societal impact. Patterns within these categories were identified, leading to the development of key themes.

Results. The findings highlight that YouTube's content often fails to align with the specific needs of this community, exposing them to irrelevant or harmful information. This results in negative consequences, including economic losses from information overload, increased domestic violence, and a gradual erosion of societal norms.

Conclusions. The study concludes that while YouTube is unsuitable for fishermen, the preference for audio-visual content underscores the need for a customised digital platform. Further research should investigate shared community spaces that combine physical and digital information, aligning with their diverse needs and capabilities to empower them with essential information.

Introduction

Accessing information through voice on smartphones has evolved from science fiction to reality, a transition first envisioned by Captain Kirk in Star Trek. This transformation has been made possible by the development of powerful smartphones and cloud-based computing. The increasing prevalence of web-enabled smartphones has created a need for quick and constant access to vital information, thereby elevating the importance of speech technology to meet users' demands for ubiquitous mobile access—any time, any place, and in any scenario (Schalkwyk et al., 2010). This emphasis on ubiquitous mobile access ensures that users can remain connected to the digital world, regardless of their location. Google, a prominent information retrieval system, is leveraging voice services to provide this ubiquitous access (Al bool et al., 2010).

Voice search enables users to input queries in spoken language and receive relevant results through system-generated transcriptions, thereby ensuring immediate and direct access to information, which is particularly popular on mobile devices (Jiang et al., 2013). Voice searches, or voice-enabled searches on mobile devices, aim to streamline the process and focus more on audio-visual content that requires less interaction with the mobile's touchscreen compared to traditional text-based inquiry and social networks (Guy, 2018; Barten, 2024). The accessibility of information using voice on smartphones is particularly beneficial for individuals with visual impairment or limited literacy skills, as it does not require visual attention or the use of hands for typing inquiries, thereby providing them with a level of independence and access to information that was previously challenging, and highlighting the potential of voice technology to empower these user groups (Dogan, 2024). In this context, the use of voice commands to access information on YouTube has significantly increased, which resulted in YouTube introducing its own voice-enabled search engine within the app (Dudharejia, n.d.). Additionally, the availability of users-made video content on YouTube (Seo & Jung, 2022) for similar types of users and the availability of important information on YouTube which is difficult to access from other sources (Lee et al., 2022) motivate users to use YouTube as a useful resource to access information.

However, the increasing sophistication of algorithmic technology presents significant obstacles for users in accessing information using voice commands on smartphones. For example, when a user is currently watching a video on YouTube, the platform automatically suggests a list of related videos (Airoldi et al., 2016). On the other hand, inaccuracies in YouTube's recognition of voice commands—caused by background noise and unclear pronunciation (Shokouhi et al., 2016) and difficulties in multilingual speech recognition vocabularies for the transcription of the videos (Schultz & Waibel, 1998; Jouvet et al., 2018) can result in the delivery of inaccurate information. While YouTube offers diverse content for entertainment, education and instruction that can teach users nearly anything (Bryant, 2020), visually impaired or illiterate individuals may encounter difficulties in effectively utilising the platform, potentially leading to serious repercussions.

Moreover, individuals who are illiterate or have limited literacy may confront substantial hurdles when using smartphones to access information, particularly when applications fail to accommodate their specific needs and preferences or do not target these individuals as potential users. These hurdles may arise from efficient access to exact information and evaluation of information. In addition, biased or one-sided suggestions within applications that offer voice command options to access information, such as the algorithms on Facebook and YouTube, can worsen these issues by promoting specific types of videos, content creators, or channels (Bessi et al., 2016). Individuals lacking a thorough understanding of these algorithms and their underlying properties may be disproportionately affected by such biases, which could contribute to the formation of ideological echo chambers among users (Kirdemir et al., 2021; Brown et al., 2022).

In this paper, we investigate the experiences of a marginalised community in a developing country, namely a small marine fishing community in Bangladesh, in accessing crucial information through smartphones and YouTube using voice commands, as well as the consequences of this practice on

their lives. In Bangladesh, people in marginalised communities predominantly use voice commands to search for multimedia content like videos, music, and occasionally weather updates on YouTube. Nevertheless, this practice can have unintended negative consequences, resulting in unfavourable outcomes despite the good intentions of those who prefer vocal commands to access information due to illiteracy. The purpose of this paper is to provide empirical evidence that can contribute to improving voice search systems by recognising the implications of voice-enabled search queries and supporting the collaborative design of targeted, user-oriented online information resource platforms. This platform would allow users to apply voice commands to access necessary information while considering their capabilities, needs, preferences, and challenges.

Related work

Voice commands in accessing information

Previous research has been conducted on the use of voice commands by different users, encompassing their performance, search pattern strategy, and effectiveness. For instance, Schalkwyk et al. (2010) comprehensively described the history of voice search efforts at Google in their research. They provided an in-depth description of the technology employed at Google and the challenges faced in making voice search a reality. Another study by Yi and Maghoul (2011) investigated the characteristics of search queries submitted from mobile devices using Yahoo! Search for Mobile. They focused on users' behaviour and found that people may use voice queries when required, which are considerably longer than typed mobile queries. Additional studies have been conducted to understand how users overcome challenges in voice user interfaces (Myers et al., 2018). When users employ voice to search for information, they frequently need to reformulate queries due to voice input errors such as speech recognition errors and improper system interruptions. In a study conducted by Jiang et al. (2013), it was found that language plays a crucial role in voice search, particularly in relation to user query reformulation behaviours. The study involved user experiments with native English speakers, revealing that these individuals often rephrase their queries using both lexical and phonetic alterations. The research sought to assess the effects of common voice input errors on users' search advancements and the efficacy of various reformulation strategies in mitigating these errors.

In another study Guy (2018) analysed voice queries in mobile web searches with a particular focus on the semantic and syntactic characteristics of these queries. His empirical evaluation revealed that the language used in voice queries more closely resembles natural language compared to typed queries. His study suggests that voice search functionality should incorporate voice-based result presentation to facilitate a fully hands-free interaction with users and also emphasise the development of new metrics for evaluating user satisfaction with voice queries. Xing et al. (2020) attempted to grasp users' performance, query characteristics, query strategy, spoken conversational search, and perception of voice search by reviewing 27 publications. They found that the majority of the studies were primarily conducted with young and well-educated populations. Moreover, Kiesel et al. (2020) applied voice-based search to investigate argument mining and conversational search. Through an online survey, they found that participants expected to primarily use voice-based argument search at home, either independently or with others. Additionally, they anticipated that such search systems would provide comprehensive information on retrieved arguments, such as the source, supporting evidence, and background knowledge on the mentioned entities or events.

Crestani and Du (2006) conducted research on the retrieval effectiveness of spoken and written queries. Their results demonstrated that using speech to express one's information needs enables a more natural expression and encourages the formulation of longer queries. However, despite spoken queries being longer in terms of the number of words, they do not appear to be significantly more effective than written ones. In a recent study, Jha et al. (2024) conducted an examination of voice-based assistants and their role in voice-based searches for accessing information on the

internet. The study observed that the methods through which users retrieve content are influenced by alterations in search engine algorithms. These modifications have implications for the digital marketing approaches employed by websites to attract visitors, establish brand identity, and execute promotional campaigns.

YouTube as a source of information

YouTube plays a significant role in entertainment by providing a vast repository of videos catering to diverse interests. It has established a unique role as a repository of popular culture, creating a diachronic archive over time and expanding synchronically in its scope (Arthurs et al., 2018). A key factor in YouTube's popularity is its accessibility; it is free to use, and while creating an account enhances the experience, it is not necessary to have an account to watch videos. In addition to entertainment, YouTube serves as an online educational platform. For example, research has shown that YouTube is utilised in Malaysia to teach social expression (Raj et al., 2019). Another study by Shoufan and Mohamed (2022) reviewed 647 papers and found that YouTube is a rich, free, easy-to-use, and enjoyable source of learning content for students. Furthermore, YouTube has been identified as an effective pedagogical tool for blending education with entertainment, creating engaging learning experiences beyond traditional classrooms (Hua, 2015). By prioritising entertainment, these channels capture larger audiences and maintain subscriber interest, making learning more accessible and enjoyable.

Moreover, YouTube is increasingly being used as a platform for disseminating health information (Madathil et al., 2015), providing valuable insights into various diseases (Delli et al., 2016; Onder et al., 2021; Diers et al., 2023). During the COVID-19 pandemic, YouTube became a significant source for accessing medical information (D'Souza et al., 2020; Andika et al., 2021). Additionally, YouTube serves as a social media platform that helps farmers improve agricultural practices (Kesherwani et al., 2022), especially through agriculture-related videos that can enhance the skills of young farmers (Paudi et al., 2022). Furthermore, YouTube serves as a valuable source of fishing techniques, enabling fishermen in Indonesia to expand their knowledge base by accessing various videos on different fishing forums (Pratiwi et al., 2023).

However, digital non-participation among marginalised populations, such as people living with disabilities, older individuals, disadvantaged youth, and women, compounds existing disadvantages (Davis & Farmer, 2016). To address this issue, several marginalised individuals with literacy challenges are using voice commands to access information via smartphones.

Information access through YouTube by individuals with low or no literacy

Several studies have explored various aspects of voice search and video content on YouTube, addressing diverse user needs across different demographics. For example, Balasuriya et al. (2018) identified barriers and opportunities in voice-activated interfaces, specifically examining how intellectually disabled users in Australia utilise voice commands to search for YouTube content. Similarly, Kversøy and Kversøy (2018) investigated how individuals with intellectual disabilities in Norway use YouTube as an alternative search tool. Research has also focused on age-related patterns of YouTube usage. Lee et al. (2022) found that older adults in South Korea actively engage with YouTube, despite concerns about fake news and costs, viewing its content as more unbiased than traditional media. Hourcade et al. (2015) analysed YouTube use among infants and toddlers, observing that interactive content, particularly educational apps, supports the development of children's perceptual and motor skills.

YouTube has also been examined regarding the personalisation and dissemination of misinformation. Hussein and Juneja (2020) demonstrated that while demographics such as age, gender, and location have minimal impact on misinformation in search results for new accounts, these factors become influential once a watch history is established. Beyond these studies, other research has addressed the broader implications of algorithmic technology, biases in video

recommendations, and factors contributing to video popularity on YouTube (Roth et al., 2020; Bryant, 2020; Foster, 2020; Lutz et al., 2021; Kirdemir et al., 2021; Gran et al., 2021; Brown et al., 2022; Haroon et al., 2022; Vybihal & Desblancs, 2023).

Despite the existing research, a significant gap remains in the literature concerning the experiences of marginalised communities in low- and middle-income countries of the Global South when using YouTube through voice commands. While prior studies primarily focus on user experiences in developed country contexts, there is limited evidence on how marginalised groups, such as small-scale marine fisherfolk in Bangladesh with limited literacy and digital skills, engage with YouTube. Specifically, it is unclear whether these communities can effectively achieve their intended goals when using the platform. Failure or difficulty in accomplishing these objectives may lead to an unjust outcome, reflecting what Beard and Langstef (2018) describe as "unjust use". To address this gap, the present study adopts an interpretive phenomenological approach to comprehensively examine why individuals from a marginalised community interact with YouTube using voice commands and the subsequent consequences of this engagement.

Data and methods

This research is part of a PhD project that explores the information cultures of a small-scale marine fishing community on a coastal island in the south-eastern part of Bangladesh, employing the information culture assessment framework by Oliver and Foscarini (2014). The project aims to identify the factors that shape the community's perceived value of information and influence their attitudes when engaging with information. The focus on this community arises from previous findings that their information environment does not adequately meet their needs (Shuva, 2017). Various barriers impede access to reliable and authentic information, including social inequality, a lack of internal information infrastructure, and educational and cultural constraints. Small-scale marine fisherfolk in Bangladesh, at the bottom of the social hierarchy, create distinct "small worlds" characterised by information filtered via their unique worldviews, social norms, and beliefs (Chatman, 1999). Despite having their traditional means of communication, the extensive use of smartphones—with limited or no awareness of responsible usage—leads them to access and use information through various applications available on these devices.

Data collection

Oliver and Foscarini (2014) define information culture as the value attributed to information and attitudes toward it. In line with this concept, the first author investigated various critical information required by fishermen to achieve different purposes in their everyday lives. He collected data on how fishermen accessed this information, focusing on their preferred methods, the reasons for these preferences, their challenges, and their level of trust in different types of information. Ethnographic fieldwork was conducted on the island between December 2023 and January 2024 to collect data for the project. This involved four focus groups, each comprising 8–10 participants and lasting approximately 1.5 hours, and 10 in-depth interviews with male and female participants. Male participants were fishermen directly involved in fishing, while female participants were family members of fishermen. This approach provided a deeper understanding of the community's information practices and the socio-cultural dynamics influencing their access to and use of critical information.

Ethnographic fieldwork

The utilisation of an ethnographic approach in social research enables a deeper understanding of behavioural factors among participants, which may not be easily distinguished through other data collection methods. In this study, field data collection took place during the fishing season, which presented challenges in recruiting participants, as most fishermen were either at sea or preparing for upcoming trips. To address this, the first author asked them when they would be available and organised the focus groups and interviews at their preferred time. Given these logistical

constraints, non-probability sampling was employed, resulting in a small but targeted sample. Snowball and purposeful sampling were utilised for focus groups and interviews, respectively, to ensure the effective selection of participants.

Embracing the ethnographic approach, the first author who conducted the fieldwork made multiple visits to various villages, including the fish landing station, where fishermen sell their catch. The author speaks the same language as the fishermen, which enabled him to interact directly with them by participating in activities such as drinking tea and sharing snacks. This immersion fostered trust and friendships, allowing the researcher to become a part of their environment. Furthermore, ethnographic fieldwork, strengthened by interviews and focus groups, responds to the realisation that verbal descriptions may not always accurately represent reality and should be complemented by observation. According to Kawulich (2005), individuals may not always act in accordance with their verbal descriptions of their behaviour. Therefore, the ethnographic approach to data collection serves to illuminate how individuals perceive and articulate their everyday experiences.

To bolster the credibility of the study findings, three key strategies were employed: prolonged engagement, reflexivity, and triangulation (Ahmed, 2024). The author's one-month stay on the island, coupled with various data collection methods, contributed to the trustworthiness of the data. The rigour and reliability of the findings were further guaranteed by the continuous application of triangulation at various phases of data collection, as outlined by Johnson et al. (2020). The incorporation of ethnographic fieldwork in conjunction with interpretive phenomenology enabled the documentation of the lived experiences of fishermen concerning their use of YouTube as a source of information.

Data analysis

To conduct data analysis, this project followed the Reflexive Thematic Analysis (RTA) approach proposed by Braun & Clarke (2006). RTA is a theoretically flexible and accessible interpretative method for qualitative data analysis that facilitates the identification and analysis of patterns or themes within a given data set (Braun & Clarke, 2012). This reflexive approach emphasises the active role of researchers in knowledge production, encouraging reflective and thoughtful engagement with both the data and the analytic process. Braun and Clarke identified several theoretical assumptions that should be addressed when conducting RTA, conceptualised as a series of continua: essentialist versus constructionist epistemologies, experiential versus critical orientation to data, inductive versus deductive analyses, and semantic versus latent coding of data (Byrne, 2021). For this research, an experiential orientation was chosen as it aligns with the objective of this research. Adopting an experiential orientation involves understanding that participants' thoughts, feelings, and experiences reflect their internal personal states, aiming to reflect the experience of a social reality. Additionally, it helps acknowledge the social construction of the research topic.

The first author was responsible for collecting data in the field, analysing it, and disseminating the findings to the other co-authors through regular meetings to ensure their familiarity with the data. The initial codes were established to categorise the types of information accessed by YouTube, including the purpose, implications of information utilisation, proficiency in using YouTube, and justification for relying on the provided information. The data was then analysed to find common underlying patterns from which themes were identified. For example, social norms were identified as an important theme because YouTube used algorithmic technology that went against the societal norms of the fishermen.

Findings

Accessing information through YouTube: weather update

Accessing timely weather updates is another crucial information for the fishermen and other residents of the island, as accurate and quickly accessible information can potentially save lives during natural hazards such as cyclones. Fishermen, in particular, frequently check weather forecasts before embarking on expeditions. Failure to get such updates can have serious repercussions. While people can get weather information via traditional sources such as television news or loudspeaker announcements from local governments and non-governmental organisations, these are not the preferred methods because updates from these methods frequently come only after the environmental conditions have already worsened. Smartphones and internet connectivity, particularly through YouTube, offer a quicker and more convenient way for fishermen to receive weather updates, even well in advance of potential cyclones in the Bay of Bengal. As a result, many fishermen consider smartphones and YouTube as their preferred methods of accessing weather information.

Although YouTube is the preferred platform for accessing weather updates, many fishermen lack the necessary skills to effectively navigate smartphone applications. As a result, they rely on the assistance of more proficient peers within their community. For instance, in one village, a fisherman with higher skills in accessing weather information via a smartphone has earned the nickname '*Abhaoa*' (meaning '*weather*' in Bengali) due to his role in helping others check weather updates on YouTube. His knowledge stems from prior experience as a volunteer in a flood monitoring project conducted by the Bangladesh Water Development Board, where he learned to use a smartphone for flood monitoring. However, like many others in the community, his illiteracy necessitates the use of voice commands to access weather information.

Accessing information through YouTube: entertainment

Fishing is a labour-intensive occupation, and due to the challenging socio-economic conditions in many fishing villages, people often begin fishing from a young age, as by engaging in fishing, they can start earning an income from childhood. On this island, becoming a fisherman is a process of learning by doing, with individuals gaining experience through repeated participation in fishing expeditions. To become a skilled fisherman, one must have prior hands-on experience, as the profession requires direct engagement with the sea, often necessitating extended periods of time away from home.

When at sea, fishermen are constantly occupied with their work, leaving them exhausted upon their return. As a result, they value their leisure time greatly and make every effort to enjoy it. Furthermore, to protect marine species, Bangladesh enforces several fishing bans throughout specific times of year, preventing fishermen from going to sea. During the fishing ban periods, entertainment becomes an important part of their everyday lives, as most fishermen lack alternative employment or skills other than fishing.

In these instances, fishermen rely on their smartphones for entertainment, with YouTube being a popular choice. They mostly watch local movies, dramas, songs, religious music, and news. Because of their illiteracy, fishermen often utilise voice commands to access content-related entertainment. However, once they are on YouTube, they frequently watch the videos recommended by YouTube and watch content nonstop without actively looking for anything in particular. Accessing YouTube is straightforward for them because it simply requires an internet connection, allowing them to quickly and easily access audio-visual content without the need to register an account. Since they tend to watch videos one after another, their internet data is consumed quickly, requiring frequent recharges, which increases their overall internet usage costs.

Consequences of using voice commands on YouTube

Impact of online gambling on marriage and family

During fieldwork, it was observed that gambling is especially popular among young fishermen, with many engaging in it regularly. Continuous interactions with the fishermen as part of the ethnographic fieldwork approach unveiled the widespread nature of this issue, which was subsequently discussed during focus groups and interviews with both male and female participants. The fishermen revealed that their introduction to gambling primarily came through YouTube advertisements, which led them to download and install gambling-specific applications on their smartphones.

One fisherman shared that they had heard of another fisherman in their village who earned a lot of money through gambling. However, most fishermen stated that they consistently lose money, and they are unsure whether the claimed earnings of another fisherman were genuine or just a rumour. Gambling among fishermen focuses on popular sports in Bangladesh, such as cricket and football. Participation typically begins with small bets deposited into gambling platforms using mobile financial services like Bkash. Initial wins often encourage participants to increase their betting amounts, resulting in significant financial losses, with some losing all of their money. During the fieldwork in one village, a casual conversation with a group of young fishermen exposed the social dynamics of gambling. As the conversation progressed, one fisherman pointed to another and shared the following story, emphasising the shared experiences and narratives that support their gambling habits.

> *This guy earned 20,000 BDT as a one-month salary from a boat owner but lost the entire amount in a single night of betting through the betting app on his smartphone. This will have implications for his home because he has a wife and children. When he fails to purchase essential groceries, as is often the case in such situations, his wife is likely to confront him, leading to tension within the family. Gambling not only creates financial strain in managing household expenses but also exacerbates domestic issues, often culminating in conflicts, household violence, and, in some cases, divorce.*

Impact of sexist content on societal norms

During a social interaction at a local tea stall, the first author requested a young fisherman demonstrate their method of accessing content related to entertainment on YouTube using voice commands. The young fisherman tried using voice commands to play a music video. However, most of the choices were movie item songs—provocative musical sequences presenting women in exposing attire adapted for male attention, some of which featured adult content. He felt embarrassed to share these suggestions with the other fishermen present, as their societal norms do not permit the consumption of such content in public spaces. The fisherman clarified that although YouTube suggests these inappropriate videos, they do not intentionally search for such content. He also mentioned that, at times, such videos appeared even without any specific requests. During this conversation, an older fisherman who was present also added—

> *in our time, we only had Jatra Pala (local stage performances) as a form of entertainment. But now, with mobile phones, the younger generation has access to content that we could never have imagined. Young fishermen earn a lot of money through fishing, which they often spend on mobile phones and internet data for entertainment. However, such provocative content accessed by the widespread use of mobile phones has contributed to social conflicts and family problems within the community.*

The concerns raised by the older fishermen were echoed by female participants during focus groups and interviews conducted as part of the fieldwork. Nearly all female respondents expressed anxiety about their young children's behaviour, particularly its deviation from accepted societal

standards. They emphasised the importance of their children upholding the same beliefs and values they hold and expressed a strong desire to prevent their children from engaging in actions contradictory to those beliefs. The women expressed significant concern over the increasing exposure to inappropriate digital content, which they associated with a rise in social issues within their community, including higher divorce rates and incidents of sexual violence, such as rape. They observed that the widespread use of smartphones and internet access had introduced young people to content that undermines established societal norms and values, fostering behavioural changes that disrupt family structures and social cohesion.

Inaccurate weather information

In one focus group, the first author asked another fisherman to demonstrate how they access weather updates using voice commands. The fisherman attempted to use YouTube but quickly realised that he did not have mobile data. The author then offered a hotspot-based mobile internet connection. The fisherman used a voice command, asking, '*Ajker Abhaoar Khobor Bolo*' (Tell today's weather update). However, the author noticed that the first video shown was from two months ago. After pointing this out, the fisherman tried again, asking, '*Ei metro Abhaohar Khobor*' (Current weather updates), but this time the video was still two days old. When the author mentioned this, another fisherman remarked that while smartphones assist in quickly accessing weather updates, they often struggle to get accurate or timely information. He added that,

> *during the recent Cyclone Midhili, we struggled to receive accurate updates. From a TV channel on YouTube, we learned that the storm was expected to make landfall at noon, but it actually arrived four hours earlier, in the early morning. As island people, getting timely and accurate weather updates is crucial for us. Otherwise, the consequences can be devastating.*

Fishermen, particularly those who are illiterate, rely heavily on the video contents of YouTube, often placing trust in live broadcasts from television channels as reliable sources of information. IPTV (Internet Protocol Television) is widely popular in Bangladesh, with many regularly watching news through these channels. However, the channel used by these fishermen provided inaccurate information for the cyclone Midhili, which resulted in the weather updates not aligning with the actual events.

Discussion

Motivation for using smartphones and YouTube

Fishermen on the island increasingly rely on smartphones and YouTube for essential information, such as weather updates and entertainment, making these tools integral to their daily routines. This dependency has cultivated habitual usage and made smartphone use more pervasive (Oulasvirta et al., 2012). For many illiterate fishermen, YouTube serves as a crucial resource, enabling them to access information through continuous video content, which compensates for their inability to read. While this reliance underscores the value of YouTube in fulfilling their informational needs, it also results in substantial economic costs to buy mobile internet due to the extensive consumption of video content. Rumpf et al. (2022) highlighted the significant economic costs associated with problematic internet usage (PUI). In this context, behaviours such as online gambling, viewing adult content, and excessive use of social media are considered forms of PUI that contribute to financial strain.

Increased internet costs due to inefficient content consumption

Fishermen, due to their inability to read, are often unable to make informed decisions about the video content they consume on YouTube. When using voice commands, they cannot understand video captions or titles, forcing them to randomly select videos. This repetitive process of checking one video after another not only increases their data usage but also contributes to the perception

that internet costs are unreasonably high. These fishermen act merely as viewers without engaging in any form of interaction on the platform, such as liking, commenting, or sharing videos (Sui et al., 2022). Moreover, they lack knowledge of data-saving techniques, such as adjusting video quality, using data-saving features, or preloading content on Wi-Fi (Whitely, 2024). Although the Internet is seen as an essential infrastructure for audio-video content access on YouTube (Lehr & Sicker, 2017), this increased consumption continues to burden them with high costs.

Problems with voice recognition for fishermen

Fishermen use voice commands in Bengali (the local language), and the system often provides outdated information, especially regarding weather updates. This happens because voice search relies on Google's audio recognition technologies, which require proper settings to be turned on (YouTube, n.d.). Unfortunately, fishermen are generally unaware of this requirement, as their smartphones were set up with the help of the local mobile technician and without proper guidance or training. Additionally, they often use YouTube without a Google account and remain unaware of the potential benefits, such as improved voice recognition technology based on saved audio data (YouTube, n.d.). This lack of user knowledge limits the effectiveness of YouTube's voice recognition system and prevents the fishermen from optimising their use of the platform to meet their informational needs.

Problems with adult content and biased recommendations

Search and recommendation algorithms shape content results based on user queries and feedback, retrieving results according to individual user preferences. These algorithms increasingly rely on user profiles to optimise the content they present (Boratto et al., 2020). However, a critical issue with these algorithms is the prevalence of popularity bias, where widely viewed or highly rated content is consistently prioritised, creating a feedback loop that amplifies its visibility over time (Kirdemir et al., 2021). In the context of fishermen seeking entertainment, such as music videos, this bias often results in the prominence of adult content and gambling advertisements. The algorithm misinterprets their repeated consumption of video content as a preference for such material, reinforcing these undesirable recommendations. Consequently, much of the online content fishermen encounter is shaped by algorithms they do not understand, leading to distorted or harmful online experiences (Lewandowsky et al., 2023).

Biases inherent in recommendation systems, stemming from historical data, can exacerbate societal inequalities by influencing everyday information access. Lack of awareness about algorithms can exacerbate the digital divide, leaving vulnerable populations at a greater disadvantage (Gran et al., 2021). When these biases touch on sensitive attributes like religion or gender, they can reinforce discrimination and deepen existing inequalities (Boratto et al., 2019). In the case of young fishermen, persistent exposure to adult content and gambling advertisements not only alters their mindset but also leads to broader societal problems, impacting their families and communities.

Information overload challenges

Due to widespread illiteracy, many fishermen are unable to read video titles, which leads to confusion and indecision when selecting content. The exponential growth of digital information has exacerbated the problem of information overload, which significantly affects decision-making and productivity (Shahrzadi et al., 2024). For these fishermen, this overload, combined with their limited literacy, makes it challenging to navigate YouTube effectively. Moreover, the algorithm-driven technology behind YouTube fails to consider their literacy, further complicating their ability to access relevant information. As a result, their access to information via YouTube becomes problematic, much like how overconsumption of junk food negatively impacts health, as it does not consider their capacity to engage meaningfully with the platform.

Conclusion and future direction

This research explored how small-scale marine fishermen, a marginalised community, use voice commands to access information on YouTube, highlighting the challenges they face. YouTube, designed to serve a broad user base, does not effectively support this community, resulting in unintended negative consequences. Importantly, fishermen are not the primary customers of YouTube; instead, they unknowingly supply their data without compensation while receiving content tailored to YouTube's actual customers—advertisers and other entities who remain anonymous. This dynamic reveals that fishermen are utilising an unaccountable platform, exposing themselves to content that is not necessarily beneficial. In this way, fishermen are exploited as tools or resources to serve external interests, leading to consuming content that may have harmful consequences. YouTube thus fails to offer the same benefits to fishermen that it provides to its target users, reflecting a lack of fairness.

The rapid technological advancements of the digital age have transformed the dissemination and accessibility of information, creating a vast online information landscape. However, marginalised communities struggle to access accurate, relevant information on digital platforms, as illustrated in this research. Information experts can address these problems by identifying the specific kinds of information sought by these communities online, as well as their preferred formats, and then co-designing digital platforms that meet their needs. Such platforms can alleviate issues like information overload by organising targeted and accessible content while minimising economic costs and mitigating the negative impacts of algorithmic technology on societal culture and values.

Future research on investigating shared community spaces that function as information grounds can also help in understanding how to enhance their effectiveness by integrating physical and digital information tailored to the diverse needs of marginalised populations. These spaces hold the potential to serve multiple purposes, including the dissemination of accurate and reliable information, while also providing opportunities for culturally aligned entertainment and social engagement. Such spaces can play a critical role in overcoming informational disparities and empowering marginalised communities to participate more fully and equitably in the digital age.

Acknowledgements

This research is part of a PhD project investigating information cultures in marginalised communities, conducted under the Participatory Research and Ownership with Technology, Information, and Change (PROTIC) II project at Monash University, Australia. The authors acknowledge the funding support from the Empowerment Charitable Trust, the participants who contributed to this research, and the staff of the ECOFISH II project of WorldFish for their logistical assistance during field data collection. The authors also extend their gratitude to the esteemed reviewers for their valuable comments.

About the authors

Joy Bhowmik is a PhD candidate in the Faculty of Information Technology at Monash University, Australia. His research interests include information culture, information empowerment of marginalised communities, transforming shared community space as information ground, and documenting the impacts of climate change. He can be contacted at joy.bhowmikk@gmail.com

Dr. Viviane Frings-Hessami is a Senior Lecturer in the Faculty of Information Technology at Monash University, Australia. Her research interests include archival theory; recordkeeping literacy; information culture; the impact of cultural and linguistic influences on the management and preservation of digital information; analogue and digital preservation, and sustainable development. She can be contacted at viviane.hessami@monash.edu.

Dr Gillian Oliver is a Professor in the Faculty of Information Technology at Monash University, Australia. Her research interests include information and data cultures. She can be contacted at gillian.oliver@monash.edu.

Dr. Md. Khalid Hossain works as a Research Fellow at the Faculty of Information Technology of Monash University, Australia. His research interests include community informatics, environmental informatics, sustainability, and ICT for Development. He can be contacted at md.khalid.hossain@monash.edu.

References

Ahmed, S. K. (2024). The pillars of trustworthiness in qualitative research. Journal of Medicine, Surgery, and Public Health, 2, 100051. https://doi.org/10.1016/j.glmedi.2024.100051

Airoldi, M., Beraldo, D., & Gandini, A. (2016). Follow the algorithm: An exploratory investigation of music on YouTube. Poetics, 57, 1-13. https://doi.org/10.1016/j.poetic.2016.05.001

Al_bool, H., Ali, M., Ibrahim, H., & Habbal, A. M. M. (2010). Hands-Free Searching Using Google Voice. *IJCSNS*, 10(8), 47.

Andika, R., Kao, C. T., Williams, C., Lee, Y. J., Al-Battah, H., & Alweis, R. (2021). YouTube as a source of information on the COVID-19 pandemic. Journal of community hospital internal medicine perspectives, 11(1), 39-41. https://doi.org/10.1080/20009666.2020.1837412

Arthurs, J., Drakopoulou, S., & Gandini, A. (2018). Researching youtube. Convergence, 24(1), 3-15.

Balasuriya, S. S., Sitbon, L., Bayor, A. A., Hoogstrate, M., & Brereton, M. (2018, December). Use of voice activated interfaces by people with intellectual disability. In Proceedings of the 30th Australian conference on computer-human interaction (pp. 102-112).

Beard, M., & Longstaff, S. (2018). *Ethical by design: principles for good technology*. Ethics Centre.

Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., ... & Quattrociocchi, W. (2016). Users polarization on Facebook and Youtube. PloS one, 11(8), e0159641. doi:10.1371/journal.pone.0159641

Boratto, L., Fenu, G., & Marras, M. (2019, April). The effect of algorithmic bias on recommender systems for massive open online courses. In European conference on information retrieval (pp. 457-472). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-15712-8_30

Boratto, L., Marras, M., Faralli, S., & Stilo, G. (2020, April). International workshop on algorithmic bias in search and recommendation (bias 2020). In European Conference on Information Retrieval (pp. 637-640). Cham: Springer International Publishing. doi: 10.1007/978-3-030-45442-5_84

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77-101.

Braun, V., & Clarke, V. (2012). *Thematic analysis*. American Psychological Association.

Brown, M. A., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users. Available at SSRN 4114905. http://dx.doi.org/10.2139/ssrn.4114905

Bryant, L.V. (2020). The YouTube Algorithm and the Alt-Right Filter Bubble. Open Information Science, 4, 85 - 90. https://doi.org/10.1515/opis-2020-0007

Byrne, D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Qual Quant* **56**, 1391–1412 (2021). https://doi.org/10.1007/s11135-021-01182-y

Chatman, E. A. (1999). A theory of life in the round. Journal of the American Society for information Science, 50(3), 207-217.

Chelba, C., & Schalkwyk, J. (2012). Empirical exploration of language modeling for the google. com query stream as applied to mobile voice search. In Mobile Speech and Advanced Natural Language Solutions (pp. 197-229). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-6018-3_8

Crestani, F., & Du, H. (2006). Written versus spoken queries: A qualitative and quantitative comparative analysis. *Journal of the American Society for Information Science and Technology*, 57(7), 881-890. https://doi.org/10.1002/asi.20350

D'Souza, R. S., D'Souza, S., Strand, N., Anderson, A., Vogt, M. N., & Olatoye, O. (2020). YouTube as a source of medical information on the novel coronavirus 2019 disease (COVID-19) pandemic. Global public health, 15(7), 935-942. https://doi.org/10.1080/17441692.2020.1761426

Davis, H., & Farmer, J. (2016, November). Digital participation: engaging diverse and marginalised communities. In Proceedings of the 28th Australian Conference on Computer-Human Interaction (pp. 672-675). http://dx.doi.org/10.1145/3010915.3011866

Delli, K., Livas, C., Vissink, A., & Spijkervet, F. K. (2016). Is YouTube useful as a source of information for Sjögren's syndrome?. Oral diseases, 22(3), 196-201.

Diers, C. S., Remvig, C., Meteran, H., Thomsen, S. F., Sigsgaard, T., Høj, S., & Meteran, H. (2023). The usefulness of YouTube videos as a source of information in asthma. Journal of Asthma, 60(4), 737-743. https://doi.org/10.1080/02770903.2022.2093218

Dogan, S. (2024). Voice Control and Voice Searching. In Future Tourism Trends Volume 2: Technology Advancement, Trends and Innovations for the Future in Tourism (pp. 159-172). Emerald Publishing Limited.

Dudharejia, M. (n.d.). How to Optimize YouTube Videos for Voice Search. E2M. Retrieved September 13, 2024, from https://www.e2msolutions.com/blog/optimize-youtube-for-voice-search/

Foster, D. (2020). Factors influencing the popularity of YouTube videos and users' decisions to watch them.

Gran, A. B., Booth, P., & Bucher, T. (2021). To be or not to be algorithm aware: a question of a new digital divide?. Information, Communication & Society, 24(12), 1779-1796. https://doi.org/10.1080/1369118X.2020.1736124

Guy, I. (2018). The characteristics of voice search: Comparing spoken with typed-in mobile web search queries. ACM Transactions on Information Systems (TOIS), 36(3), 1-28. https://doi.org/10.1145/3182163

Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z., & Wojcieszak, M. (2022). Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations. arXiv preprint arXiv:2203.10666.

Hourcade, J. P., Mascher, S. L., Wu, D., & Pantoja, L. (2015, April). Look, my baby is using an iPad! An analysis of YouTube videos of infants and toddlers using tablets. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 1915-1924).

http://dx.doi.org/10.1145/2702123.2702266 Families and Their Use of Technology CHI 2015, Crossings, Seoul, Korea

Hua, K. (2015, June 23). Education as Entertainment: YouTube Sensations Teaching The Future. *Forbes*. https://www.forbes.com/sites/karenhua/2015/06/23/education-as-entertainment-youtube-sensations-teaching-the-future/?sh=54f09f247c2e

Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring misinformation in video search platforms: An audit study on YouTube. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), 1-27. https://doi.org/10.1145/3392854

Jha, A. K., Singhal, N., & Chhabra, A. (2024). Analysis of Voice-Based Searches and How They Affect Digital Marketing. Educational Administration: Theory and Practice, 30(5), 13914-13921. Doi: 10.53555/kuey.v30i5.6119

Jiang, J., Jeng, W., & He, D. (2013). How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 143-152). https://doi.org/10.1145/3182163

Johnson, J. L., Adkins, D., & Chauvin, S. (2020). A review of the quality indicators of rigor in qualitative research. American journal of pharmaceutical education, 84(1), 7120. https://doi.org/10.5688/ajpe7120

Jouvet, D., Langlois, D., Menacer, M. A., Fohr, D., Mella, O., & Smaïli, K. (2018). Adaptation of speech recognition vocabularies for improved transcription of youtube videos. Journal of International Science and General Applications, 1(1), 1-9.

Kawulich, B. B. (2005, May). Participant observation as a data collection method. In *Forum qualitative sozialforschung/forum: Qualitative social research* (Vol. 6, No. 2).

Kesherwani, Bhagyashree., Rout, Sandeep., Padhy, Deepayan., & Ravichandran, S. (2022). Social media help farmers: For improving agriculture practices. International Journal of Biology, Pharmacy and Allied Sciences, 11(1). https://doi.org/10.31032/IJBPAS/2022/11.1.1051

Kiesel, J., Lang, K., Wachsmuth, H., Hornecker, E., & Stein, B. (2020, March). Investigating expectations for voice-based and conversational argument search on the web. In Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (pp. 53-62). https://doi.org/10.1145/3343413.3377978

Kirdemir, B., Kready, J., Mead, E., Hussain, M. N., & Agarwal, N. (2021, April). Examining video recommendation bias on YouTube. In International Workshop on Algorithmic Bias in Search and Recommendation (pp. 106-116). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-78818-6_10

Kversøy, K. S., & Kversøy, S. D. (2018). Exploring career possibilities for people with intellectual disabilities. Indian Journal of Career and Livelihood Planning, 7(1), 39-47.

Lee, J., Gwak, S., Gwon, J., Park, J., Eom, S., Hong, S., ... & Jung, H. (2023). Exploring the community of older adult viewers on YouTube. Universal Access in the Information Society, 22(4), 1393-1404. https://doi.org/10.1007/s10209-022-00918-3

Lehr, W., & Sicker, D. (2017). Would you like your Internet with or without video?. U. Ill. JL Tech. & Pol'y, 73.

Lewandowsky, S., Robertson, R. E., & DiResta, R. (2023). Challenges in understanding human-algorithm entanglement during online information consumption. Perspectives on Psychological Science, 17456916231180809. https://doi.org/10.1177/17456916231180809

Lutz, M., Gadaginmath, S., Vairavan, N., & Mui, P. (2021, December). Examining political bias within youtube search and recommendation algorithms. In 2021 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1-7). IEEE.

Martijn, B. (2024, June 1). How search travel industry: How can your company benefit? www.revfine.com. Retrieved September 11, 2024, from https://www.revfine.com/voice-search-travel-industry/

Myers, C., Furqan, A., Nebolsky, J., Caro, K., & Zhu, J. (2018, April). Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-7). https://doi.org/10.1145/3173574.3173580

Oliver, G., & Foscarini, F. (2014). *Records management and information culture: Tackling the people problem*. Facet Publishing.

Onder, M. E., & Zengin, O. (2021). YouTube as a source of information on gout: a quality analysis. Rheumatology International, 41(7), 1321-1328. https://doi.org/10.1007/s00296-021-04813-7

Oulasvirta, A., Rattenbury, T., Ma, L., & Raita, E. (2012). Habits make smartphone use more pervasive. Personal and Ubiquitous computing, 16, 105-114. DOI 10.1007/s00779-011-0412-2

Paudi, M. H., Din, R., & Othman, N. (2022). Agricultural Video as A Learning Medium for Young Farmer. International Journal of Academic Research in Progressive Education and Development, 11 (1), 729-742. http://dx.doi.org/10.6007/IJARPED/v11-i1/12142

Pratiwi, M. A., Wahyu, H. R., & Normande, B. E. (2023). Understanding knowledge acquisition, adaptive strategies, challenges, and preservation methods among traditional fishermen in a digital age. In BIO Web of Conferences (Vol. 70, p. 05008). EDP Sciences. https://doi.org/10.1051/bioconf/20237005008

Raj, S. A. S., Ann, W. H. T., Subramaniam, P. A., & Yunus, M. M. (2019). Using YouTube as a platform to learn social expression. Creative Education, 10(2), 288-296. DOI: 10.4236/ce.2019.102023

Roth C, Mazières A, Menezes T (2020) Tubes and bubbles topological confinement of YouTube recommendations. PLoS ONE 15(4): e0231703. https://doi.org/10.1371/journal.pone.0231703

Rumpf, H. J., Effertz, T., & Montag, C. (2022). The cost burden of problematic internet usage. *Current Opinion in Behavioral Sciences*, 44, 101107.

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., ... & Strope, B. (2010). "Your word is my command": Google search by voice: A case study. *Advances in Speech Recognition: Mobile Environments, Call Centers, and Clinics*, 61-90.

Schultz, T., & Waibel, A. (1998, February). Multilingual and crosslingual speech recognition. In Proc. DARPA Workshop on Broadcast News Transcription and Understanding (pp. 259-262).

Seo, W., & Jung, H. (2022). Challenges and opportunities to improve the accessibility of YouTube for people with visual impairments as content creators. Universal Access in the Information Society, 21(3), 767-770. https://doi.org/10.1007/s10209-020-00787-8

Shahrzadi, L., Mansouri, A., Alavi, M., & Shabani, A. (2024). Causes, consequences, and strategies to deal with information overload: a scoping review. *International Journal of Information Management Data Insights*, 4(2), 100261.

Shokouhi, M., Ozertem, U., & Craswell, N. (2016, April). Did you say U2 or YouTube? Inferring implicit transcripts from voice search logs. In Proceedings of the 25th International Conference on World Wide Web (pp. 1215-1224).

Shoufan, A., & Mohamed, F. (2022). YouTube and education: A scoping review. IEEE Access, 10, 125576-125599.

Shuva, N. Z. (2017). The information practices of the fishermen in the Bay of Bengal, Bangladesh. *Open Information Science*, 1(1), 21-40. DOI 10.1515/opis-2017-0003

Sui, W., Sui, A., & Rhodes, R. E. (2022). What to watch: Practical considerations and strategies for using YouTube for research. Digital Health, 8, 20552076221123707. DOI: 10.1177/20552076221123707

Vybihal, J., & Desblancs, M. (2023). Analysis of the Impact of Algorithms on Siloing Users: Special Focus on YouTube. In AI and Society (pp. 51-69). Chapman and Hall/CRC.

Whitley, A. (2024, January 17). How Much Data Does YouTube Use? airalo. Retrieved September 14, 2024, from https://www.airalo.com/blog/how-much-data-does-youtube-use

Yi, J., & Maghoul, F. (2011, March). Mobile search pattern evolution: the trend and the impact of voice queries. In Proceedings of the 20th international conference companion on World wide web (pp. 165-166).

YouTube. (n.d.). Find videos with Voice Search. YouTube Help. https://support.google.com/youtube/answer/9372650?hl=en&co=GENIE.Platform%3DAndroid

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Beyond language archives: proposing the archival community informatics framework as an interdisciplinary link to revitalization lexicography

Merrion Dale Frederick, Ana Roeschley, and Oksana L. Zavalina

DOI: https://doi.org/10.47989/ir30iConf46942

Abstract

Introduction. Lexicography, or the practice of compiling dictionaries, has profound impacts on how information, culture, and identity are understood and communicated. The emergence of revitalization lexicography serves as an example of community informatics practices in which societal implications of dictionary building are investigated and refined by and for minority language communities to reverse epistemic and cultural erasure.

Method. To advance engagement of information and archival sciences with revitalization lexicography, this conceptual paper proposes the archival community informatics framework. The framework is then mapped to current revitalization lexicography projects as an entry point to understanding the information practices of communities engaging in revitalization efforts.

Results. Mapping the framework to existing initiatives reveals commonalities between archival and lexicographical community informatics, as well as areas for methodological support from the information science and archives fields to aid revitalization lexicographers in the cultural heritage preservation functions of minority language dictionaries.

Analysis. This paper argues for further information science analysis of dictionaries—with a conceptualization of dictionaries as archival repositories—housing collections of word definitions, cultural values, traditions, and information practices. The authors posit that through interdisciplinary engagement, the information needs of minority language community members and revitalization lexicographers can be supported.

Introduction

Studies of language archives—repositories housing language data, including linguistic fieldnotes and grammatical sketches among other materials—have contributed to the establishment of disciplinary ties between information science (IS) and linguistics (Burke et al., 2020). Other products of linguistic practice, like dictionaries, are often excluded from the IS inquiry, beyond studies of their functions as materials within collections. This paper argues for IS analysis of dictionaries as archival repositories which house collections of word definitions, cultural values, traditions, and information practices. To encourage IS and archival engagement with lexicography, this conceptual paper proposes the archival community informatics (ACI) Framework of archival functions in *revitalization lexicography*—the practice of constructing dictionaries to revive a minority language (ML) while prioritizing the input and needs of ML community members (Anderson, 2020). The framework is then mapped to current revitalization projects as an entry point to conceptualizing the information practices of ML community members and lexicographers engaging in revitalization efforts.

The archival lives of dictionaries

Community informatics (CI) refers to applied research efforts that promote the use and development of information technologies to benefit communities (Averweg & Leaning, 2011). A defining feature of CI is that uses of information technologies are guided by a '*proactive social responsibility*' to people instead of pursuits of information or technological advancements for their own sakes (Gilliland & Flinn, 2013, p. 20). Thus, in archival community informatics (ACI), the cultural and information needs of (often underserved) communities are prioritized above other archival functions such as custodial control and preservation of archival holdings (Bastian & Alexander, 2009; Caswell et al., 2018; Lee, 2021). While preservation and care for archival holdings are key to ACI, care for materials never overrides care for people. Rather, the archival stewardship of materials is done in service of communities in community-based archives. Cook's (2013) treatise on archival paradigms describes this as part of the community paradigm of archives, wherein archivists '*work in the community to encourage archiving as a participatory process shared with many in society*' (p. 114).

Furthermore, ACI work explicitly seeks to center the needs of individuals who are subjects or creators of archival materials, rather than outsider researchers (Bastian, 2003; Carter, 2017; Roeschley, 2023). Similarly, when engaging in revitalization lexicography, lexicographers are required to shift their dictionary building practices away from outsider linguistic researchers to focus on ML community members whose familial languages are suppressed and endangered. Through this community-centered approach to lexicography, conceptualizations of dictionaries shift from technical reference resources to repositories of community cultural heritage.

Of the thousands of world languages, the majority are incredibly localized and often under-resourced, with 41% classified as endangered (The Language Conservancy, 2023). Language endangerment is exacerbated by cultural marginalization of ML communities, political instability, and natural disasters—factors that also increase loss of life and forced migration of language communities from their homelands (Woodbury, n.d.; Chelliah, 2021). Historically, the eradication of Indigenous languages and cultures by colonial authorities was common (Mojela, 2008; Pine & Turin; 2017, Hinton et al.; 2018, Anderson, 2020). In cases where Indigenous and other MLs were preserved by dominant groups and governments, it was often with aims for the languages '*to transmit and disseminate an imagined Christian modernity*' (Pine & Turin, 2017, p. 3).

While dictionaries were purported to be value neutral, lexicography played a central role in epistemic and cultural erasure. Through the Enlightenment Era, dictionaries served larger '*nation-building*' efforts toward linguistic purity, with majority European languages positioned as '*gold standards*' and instruments of '*a power that academies seized to relegate minority languages or*

dialects to a secondary position' (de Castro Fario Salgado, 2021, p. 67). Remnants of epistemic and cultural annihilation continue as even dictionaries created for the preservation of MLs have historically been subjected to institutional/colonialist objectives as opposed to the specific needs, desires, and overall flourishing of the language communities in question (Sear & Turin, 2021).

Efforts to reverse these practices involve language revitalization and reclamation projects with overarching goals of supporting initiatives that result in young people in ML communities understanding and speaking their heritage languages. While such efforts can only succeed if fueled by the ML community, they can often benefit from collaboration and knowledge-sharing with outside researchers. The term *revitalization lexicography*, coined by Anderson (2020) for the Tunica Language Project, positions dictionary-making as a CI tool to reverse ML loss with,

> *the explicit acknowledgment that collecting, cataloging, and organizing language is not a neutral act and that the way in which said lexicography is conducted has a material impact on language use, perceptions, and attitudes* (p. 7).

This conceptualization of lexicography breaks from Western traditions which prioritize clinical linguistic inquiry over the well-being of language community members. Rather, it is equivalent to conceptualizations of archival practice stemming out of critical archival studies which also break from scholarly and professional traditions regarding the stewardship of archival materials (Caswell et al., 2017; Wick, 2017; Ghaddar, 2022). Often grounded in ACI, critical archival scholars argue that prescribed depictions of archival practice as neutral, standardized, and purely technical work allow archivists to hide their '*social and cultural responsibilities*' in constructing the archival record (Jimerson, 2006, p. 28).

Just as revitalization lexicography places importance on language communities over linguistic researchers, community archives prioritize '*community values over the centralized and bureaucratized control and ownership of traditional mainstream archives*' (Poole, 2020, p. 662). Within community archives that center and serve the needs of minoritized and underrepresented communities, archivists and other community memory workers reject neutrality and

> *value people over stuff. The stuff—what gets collected—is only as important as it enables connections between people, who use the stuff to share stories, transmit memory, and build relationships* (Caswell, 2021, p. 18).

Similarly, the CI work in revitalization lexicography prioritizes community needs over technical and pure linguistic research functions (Cablitz, 2011). Because language is so intertwined with identity, a dictionary can occupy overlapping linguistic and extralinguistic '*social roles*' including (but not limited to) a community pedagogical development tool, an access point for certain cultural knowledge, a reference material for language learners, an instrument to enhance language prestige, and, in some cases, a source of grammatical information for outside linguists in addition to the intended language community (Anderson, 2020; Chelliah, 2021; Chelliah, 2023).

Despite these obvious overlaps, there is little existing scholarship connecting the community-based approaches in archival (and wider IS) and lexicographical fields. The presence of disciplinary silos is especially concerning as it suggests that there is little to no methodological support from the IS and archives fields to aid revitalization lexicographers in implementing and stewarding the cultural heritage preservation functions of ML dictionaries. Increasing interdisciplinarity between IS and linguistics has resulted in language archives better serving the needs of their users—including users who belong to ML communities (Burke, 2023; Zavalina & Chelliah, 2022). However, dictionaries are not traditionally thought of as archival repositories, but only as materials sometimes found in an archive. Could applying a functional archival lens on dictionaries and lexicography aid in the ACI work done by revitalization lexicographers?

Proposing a framework of archival community informatics functions present in revitalization lexicography

A previous scoping review investigation resulted in the ACI framework (presented in Figure 1), which identifies ACI functions that are present in revitalization lexicography (Frederick, 2024). This conceptual paper further tests this ACI framework by mapping its functions to three revitalization lexicography initiatives, with the aim of investigating functional areas where IS and archival studies fields can support cultural preservation and CI efforts in the revitalization lexicography field.

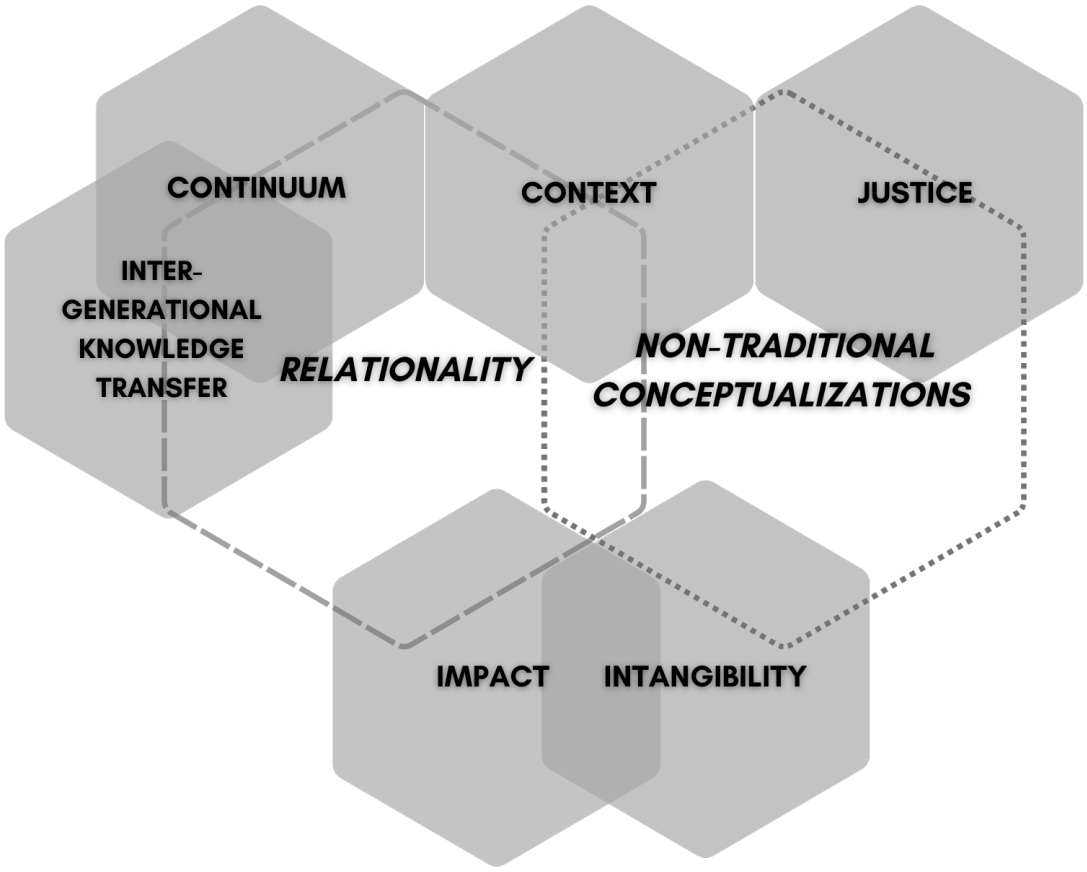


Figure 1. Archival community informatics framework of archival functions in revitalization lexicography

As shown in Figure 1, ACI functions in revitalization lexicography are interrelated, and much like archives and dictionaries, do not exist in a vacuum. Rather, these functions are most useful for CI research when applied in conjunction with each other. The foundational level includes two overlapping foundational root concepts: *relationality* and *non-traditional conceptualizations*. *Relationality* illustrates how memory, evidence, people, and place are all interrelated within their own societal contexts. As such, they reveal both power structures and the ties that bind people to each other and to their epistemic values. *Non-traditional conceptualizations* point to information behaviours, preservation practices, and theory-development that are outside of the traditional Western canon. Flowing out of *relationality* and *non-traditional conceptualizations* are six functions of ACI which are overwhelmingly applicable to revitalization lexicography: *impact*, *inter-generational knowledge transfer*, *continuum*, *context*, and *justice*, and *intangibility* (described in Table 1).

Foundational Root	Function Name	Definition
Relationality	*Impact* (overlaps *Intangibility*)	Umbrella concept for certain, often immediately intangible, embodied notions such as affect, identity, collective memory. The impact in question can be positive or negative depending on the broader circumstances.
	Inter-generational Knowledge Transfer (overlaps *Continuum*)	Details the act or desire to pass one's knowledge and/or heritage on to one's younger generation for the sake of cultural continuity and healing generational trauma.
	Continuum (overlaps *Inter-generational Knowledge Transfer*)	Broad label for ACI practices that adhere to records continuum principles such as collaborative reminiscence, multiple provenances, multiple epistemologies (including participatory methodology and other co-creator archiving scenarios), fluidity, and semantic stability.
	Context (stems from both foundational roots)	Minority/marginalized/Indigenous perspectives prioritized, asserted, and reclaimed to contextualizing narratives, information, and functions of CI work.
Non-traditional Conceptualization	*Justice*	Anti-colonial preservation practices, with the repository playing a part in the pursuit of justice for community members to meet their personal memory, identity, and accountability needs.
	Intangibility (overlaps *Impact*)	Cultural heritage, orality, dance, storytelling, performance, embodiment, and Indigenous ways of knowing.

Table 1. Archival community informatics functions applicable to revitalization lexicography

Applying the archival community informatics framework to revitalization lexicography projects

Just as community memory workers, archivists and archival scholars have worked to reconceptualize provenance to center the colonized rather than the colonizer, members of language communities and linguists are moving away from methodologies that serve academic researchers to ones that serve language and memory needs of ML communities. Three revitalization initiatives are described and then mapped onto the framework components. Chosen purposefully to illustrate the differences between projects, the individual initiatives vary in design, language community, and institutional capacity. However, they all center their language communities in their work.

The New Tunica Dictionary (NTD), under the umbrella of Patricia Anderson's involvement with the Tunica Language Project, is a community-driven research initiative affiliated with the Tunica Biloxi Tribe which prioritizes the preservation of the Tunica language spoken in Northeastern Louisiana and Southwestern Mississippi (Anderson, 2020). The Jejueo Talking Dictionary (JTD) is a collaborative digital resource designed for use in revitalization of the Jejueo language indigenous to the South Korean Jeju Island. Arguably the most comparable to a digital library or cultural heritage archive, JTD integrates '*existing annotated video corpora of Jejueo songs, conversational genres and regional mythology into a multimedia database*', and is positioned by its creator, linguist Moira Saltzman (2017), as a '*repository for oral history*' that will ideally be used in the context of Jejueo language acquisition (p. 122). Finally, the LearnCajun (LC) dictionary app is an independent community project started by Luke Romero, a non-linguist native of St. Martinville Parish and heritage speaker of Louisiana Regional French. Like many independent community archives that

do not always follow prescriptive archival practices, LC does not conform to traditional linguistic approaches and emphasizes the cultural heritage dimensions of the language.

Through a qualitative content analysis approach (Spurgin & Wildemuth, 2016), the research team analyzed each initiative's dictionary, website, and related documentation (including publications about the efforts) for presence of the framework concepts. Each project is uniquely built to serve the needs of its ML community. As such, they look and function in disparate ways. However, in analyzing their CI work, the research team was able to map aspects of each to the functions within the ACI Framework as can be seen in Figure 2.

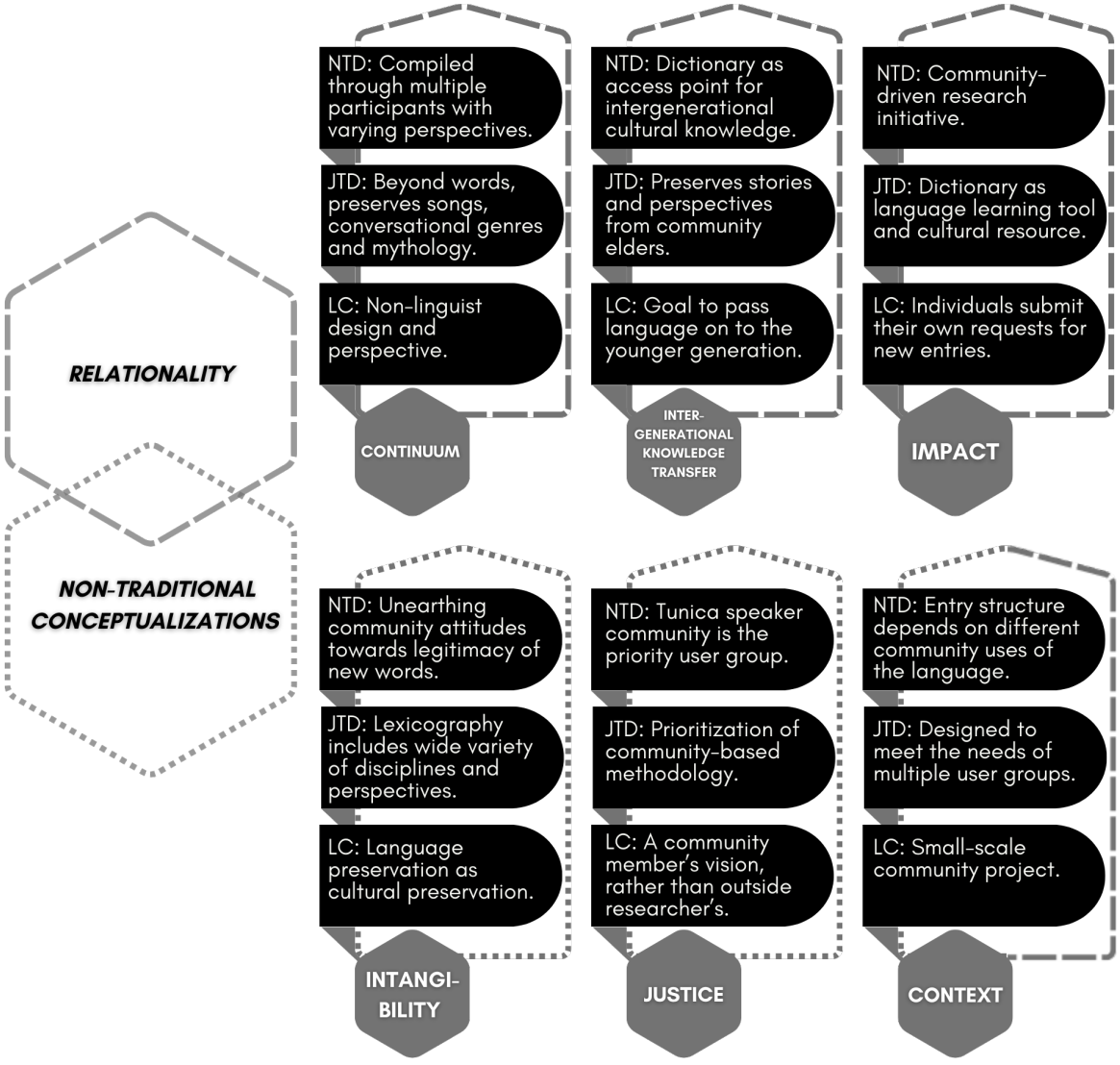


Figure 2. Applications of the archival community informatics framework for revitalization lexicography concepts in NTD, JTD, and LC

Discussion

Harris (2009) observed the '*many lenses available to archivists in their reading of archive [and that] lenses from outside archival discourses offer abundant riches*' (p. 141). In ACI, working actively with (non-archivist) priority stakeholders can result in an ontological shift transforming archives from static document repositories to living embodiments of memory (Lee, 2016; Roeschley & Kim, 2019).

In revitalization lexicography, needs of ML community members can be prioritized through adherence to the archiving *continuum* principles of multiple epistemologies and participatory/community-based workflows. Saltzman (2017) has appealed to *continuum* concepts by including a very diverse and interdisciplinary group of experts in the Jejueo Talking Dictionary Project for a '*repository for oral history*' (p. 122) to meet the needs of multiple user groups and operate as both a language learning tool and a cultural preservation resource.

There is often a clear intention for *intergenerational knowledge transfer* through both archival and lexicographic work (Wallace et. al, 2014; Saltzman, 2017; Dallwitz et al., 2023; Debenport et al., 2023). Saltzman (2017) observes that unlike with community-made cultural heritage materials in the dictionary, many members of the Jejueo language community are unmotivated to use materials designed for academic researchers. Similarly, while LearnCajun does not conform to all prescribed linguistic practices, it operates with total community autonomy and involves community elders in the creation of dictionary entries.

Context can hold a spectrum of (dis)empowering implications in the archival sphere (O'Toole, 2002; Bhebhe & Ngoepe, 2021). As seen in NTD, JTD, and LC projects, CI practices utilize context-based decision making -- rather than following one-size-fits-all prescriptive approaches -- resulting in dictionaries fulfilling multiple cultural heritage, linguistic, and information needs. *Context* also drives deep engagement with communities in the construction of both dictionaries and community archives. Thieberger (2015) notes that '*the greater depth of knowledge of a "speaker as lexicographer" will allow a dictionary to explore idiomatic and poetic uses of the language that may escape the outsider lexicographer*' (p. 5, emphasis added), emphasizing that a dictionary's usability for the language community is exponentially enhanced by prioritizing the language community narrative (as opposed to catering to linguists as the primary user group).

From an interdisciplinary perspective merging lexicography with archiving principles, the '*idiomatic and poetic uses*' (Thieberger, 2015, p.5) could also be regarded as manifestations of a language community's *intangible* cultural heritage with connections to Indigenous ways of knowing, orality, and embodiment described in archival literature (Sutherland, 2016; Kutay, 2019; Egan, 2022). Notions of *context*, *intangibility*, and *justice* emphasize the importance of the language community's autonomy over what content is included/excluded, how it is formatted, and who is allowed access. A notable example from the NTD is the ongoing discussion gauging community attitudes towards the inclusion of new words (neologisms) in the dictionary (Anderson, 2020).

Archival and lexicographic rejections of traditional formulaic protocols, as well as investments in relationship building allow for *justice* and liberatory consequences—with a marked *impact* on the communities involved. Calls for an integration of critical theory in lexicographic and archival spheres examine the implications that (de)colonial epistemologies have on collaborative repository building (Leonard, 2017; McKemmish, 2019; Sear & Turin, 2021). In terms of *impact*, the creation of a dictionary, particularly for an under-resourced ML, necessarily involves lexicographers (community members or otherwise) taking stances on linguistic features, which can be polarizing to a language community. It is also significant to note the ways that identity feeds into the concept of *impact* in *relation* to community, dictionaries, and other memory work (MacNeil, 2018). As such, these manifestations of *impact* and justice directly tie back to Tsunoda's (2006) observation that language reclamation efforts can positively enhance the language ideologies of a speaker community—working to transform imposed majority language prestige into connection with, and appreciation of ML heritage.

Revitalization lexicography is a direct answer to epistemic injustice. While a linguistic discipline, lexicography is also an archiving practice. Imposed supremacy values of majority language groups are preserved in dictionaries—as are cultural traditions of ML communities. While mapping the ACI framework to existing revitalization initiatives reveals common functions in the construction

of archives and dictionaries, further IS research on lexicography is needed. Future steps in this project include interviews and focus groups with revitalization lexicographers and ML community members to better understand their experiences, their information needs, and areas of interdisciplinary collaboration.

Conclusion

Dictionaries preserve more than words—they embody cultural heritage and epistemic values. Like archives, they are selectively constructed to both amplify and silence. By embracing revitalization lexicography as a type of CI practice, IS and archival scholars have an opportunity to serve revitalization efforts through methodological support for heritage preservation workflows, information needs research, dictionary user support, and the many other functions of ML dictionaries. Continuing explorations into the ways archival functions operate in the domain of lexicography can result in transformative spaces for stewarding CI and language preservation.

About the authors

Merrion Frederick is a Ph.D. candidate in Department Information Science, University of North Texas, USA. Her research interests focus on the intersection between archival theory/praxis and community-based lexicography for endangered languages. She can be contacted at JessicaDale@my.unt.edu

Ana Roeschley is Assistant Professor and Director of Archival Studies in Department of Information Science, University of North Texas, USA. She received her Ph.D. from University of North Texas, and her research interests are in developing improved archival and information services to under-served individuals and communities, including members of minority language communities. Dr. Roeschley can be contacted at Ana.Roeschley@unt.edu

Oksana L. Zavalina is Professor in department of Information Science, University of North Texas, USA. She received her Ph.D. from University of Illinois, and her research interests are in the information organization and user needs in libraries and other repositories, including language archives and community archives. Dr. Zavalina can be contacted at Oksana.Zavalina@unt.edu

References

Anderson, P. (2020). *Revitalization lexicography: The making of the New Tunica Dictionary*. University of Arizona Press. http://ebookcentral.proquest.com/lib/unt/detail.action?docID=6380629

Averweg, U. R., & Leaning, M. A. (2011). Visions of community: community informatics and the contested nature of a polysemic term for a progressive discipline. *Information Technologies & International Development*, 7(2), 17-29. https://itidjournal.org/index.php/itid/article/view/732/312.html

Bastian, J. A. (2003). Owning memory: How a Caribbean community lost its archives and found its history. Westport, Connecticut: Libraries Unlimited.

Bastian, J. A., & Alexander, B. (Eds.). (2009). Community archives: The shaping of memory. London, England: Facet Publishing.

Bhebhe, S., & Ngoepe, M. (2021). Elitism in critical emancipatory paradigm: National archival oral history collection in Zimbabwe and South Africa. *Archival Science*, *21*, 155-172. https://doi.org/10.1007/s10502-020-09351-y

Burke, M., Zavalina, O.L., Phillips, M.E., & Chelliah, S. (2020). Organization of knowledge and information in digital archives of language materials. *Journal of Library Metadata*, 20(4), 185-217 [11442 words]. https://doi.org/10.1080/19386389.2020.1908651

Burke, M. (2023). *Designing Archival Collections to Support Language Revitalization: Case Study of the Boro Language Resource.* [Unpublished doctoral dissertation]. University of North Texas.

Cablitz, G. H. (2011). Documenting cultural knowledge in dictionaries of endangered languages. *International Journal of Lexicography*, *24*(4), 446–462. https://doi.org/10.1093/ijl/ecr017

Carter, E. (2017). "Setting the record straight": The creation and curation of archives by activist communities. A case study of activist responses to the regeneration of Elephant and Castle, South London. Archives and Records, 38(1), 27–44. https://doi.rg/10.1080/23257962.2016.1260532

Caswell, M., Punzalan, R., & Sangwand, T. K. (2017). Critical archival studies: An introduction. *Journal of Critical Library and Information Studies*, 1(2). https://doi.org/10.24242/jclis.v1i2.50

Caswell, M., Gabiola, J., Zavala, J., Brilmyer, G., & Cifor, M. (2018). Imagining transformative spaces: The personal–political sites of community archives. Archival Science, 18(1), 73–93. https://doi.org/10.1007/s10502-018-9286-7

Caswell, M. (2021). *Urgent Archives: Enacting Liberatory Memory Work.* Routledge.

Chelliah, S. L. (2021). *Why language documentation matters*. Springer International Publishing AG. http://ebookcentral.proquest.com/lib/unt/detail.action?docID=6462444

Chelliah, S. L. (2023). Creating dictionaries together. [Unpublished manuscript]. University of Indiana – Bloomington.

Cook, T. (2013). Evidence, memory, identity, and community: Four shifting archival paradigms. *Archival Science*, 13(2–3), 95–120. http://doi.org/10.1007/s10502-012-9180-7

Debenport, E., Goeman, M., Montenegro, M., & Wynne, M. (2023). How a dictionary became an archive: Community language reclamation using the Mukurtu Content Management System. *Dictionaries: Journal of the Dictionary Society of North America*, 44(2), 29-55. https://doi.org/10.1353/dic.2023.a915064

de Castro Fario Salgado, A. M. (2021). *Terminological Methods in Lexicography: Conceptualising, Organising, and Encodeing Terms in General Language Dictionaries.* [Doctoral dissertation, Universidade Nova de Lisboa]. Repositorio Universidade NOVA. http://hdl.handle.net/10362/137023

Egan, P. (2022). In search of the item: Irish traditional music archived fieldwork and the digital. *Archival Science* (23), 45-63. https://doi.org/10.1007/s10502-021-09382-z

Frederick, M. D. (2024, October). A Case Study Applying a Framework of Contemporary Archival Theory to Lexicography Projects for Under-Resourced Languages. In *Proceedings of the ALISE Annual Conference*.

Ghaddar, J. J. (2022). Provenance in place: crafting the Vienna Convention for global decolonization and archival repatriation. In *Disputed Archival Heritage* (pp. 49-86). Routledge.

Gilliland, A., & Flinn, A. (2013, October). Community archives: What are we really talking about. In *CIRN Prato Community Informatics Conference.*

Harris, V. (2009). Against the grain: Psychologies and politics of secrecy. *Archival Science*, 9, 133-142. https://doi.org/10.1007/s10502-009-9096-z

Hinton, L., Huss, L. & Roche, G. (Eds.). (2018). *The Routledge handbook of language revitalization.* Routledge.

Jimerson, R. (2006). Embracing the power of archives. *The American Archivist*, 69(1), 19-32.

Kutay, C. (2019). Locating knowledge. *Archives & Manuscripts*, 47(1), 72-93. http://doi.org/10.1080/01576895.2018.1551143

Lee, J. A. (2016). Be/longing in the archival body: Eros and the 'endearing' value of material lives. *Archival Science*, 16, 33-51. http://doi.org/10.1007/s10502-016-9264-x

Lee, J. A. (2021). Archives as spaces of radical hospitality. *Australian Feminist Studies*, 36 (108) 156-164. https://doi.org/10.1080/08164649.2021.1969520

Leonard, W. Y. (2017). Producing language reclamation by decolonizing 'Language.' *Language Documentation and Description*, 14(16), 15-36. https://www.lddjournal.org/article/1179/galley/2424/download/

MacNeil, H., Duff, W., Dotiwalla, A., & Zuchniak, K. (2018). 'If there are no records, there is no narrative': The social justice impact of records of Scottish care-leavers. *Archival Science*,18, 1-28. https://doi.org/10.1007/s10502-017-9283-2

McKemmish, S., Bone, J., Evans, J., Golding, F., Lewis, A., Rolan, G., Thorpe, K., & Wilson, J. (2019). Decolonizing recordkeeping and in childhood out-of-home care and Indigenous archival collections. *Archival Science*, 20, 21-49. https://doi.org/10.1007/s10502-019-09321-z

O'Toole, J. (2002). Cortes's notary: Thy symbolic power of records. *Archival Science*, 2, 45-61. https://doi.org/10.1007/BF02435630

Pine, A., & Turin, M. (2017). Language revitalization. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics*. Oxford University Press. http://doi.org/10.1093/acrefore/9780199384655.013.8

Poole, A. H. (2020). The information work of community archives: a systematic literature review. *Journal of Documentation*, 76(3), 657-687. https://doi.org/10.1108/JD-07-2019-0140

Roeschley, A., & Kim, J. (2019). 'Something that feels like a community': The role of personal stories in building community-based participatory archives. *Archival Science*, 19, 27-49. https://doi.org/10.1007/s10502-019-09302-2

Roeschley, A. (2023). 'They care enough to document people's stories': Using ethnographic methods to understand collection day outreach events in participatory archives. *Library & Information Science Research*, 45(2), 101234. https://doi.org/10.1016/j.lisr.2023.101234

Saltzman, M. (2017). A collaborative online database for language revitalization. In *Proceedings of the Second Workshop on the Use of Computational Methods in the Study of Endangered Languages.* Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-0117

Sear, V., & Turin, M. (2021). Locating criticality in the lexicography of historically marginalized languages. *History of Humanities*, 6(1), 237-259. https://doi.org/10.1086/713266

Spurgin, K. M., & Wildemuth, B. M. (2016). Content analysis. *Applications of social research methods to questions in information and library science*, 307-317.

Sutherland, T. (2016). From (archival) page to (virtual) stage: The virtual vaudeville prototype. *American Archivist*, 79(2), 392-416. https://doi.org/10.17723/0360-9081-79.2.392

The Language Conservancy. (2023). Yavapai-Apache get federal grant to revitalize languages. https://languageconservancy.org/yavapai-apache-get-federal-grant-to-revitalize-languages-journal-az/

Tsunoda, T. (2006). *Language endangerment and language revitalization: An introduction diagnosis, factors, language attitude*. Walter de Gruyter.

Wallace, D. A., Pasick, P., Bermen, Z., & Weber, E. (2014). Stories for Hope, Rwanda: A psychological, archival collaboration to promote healing and cultural continuity through intergenerational dialogue. *Archival Science*, 14, 275-306. http://doi.org/10.1007/s10502-014-9232-2

Wick, A. (2017). We're all vegans here: The twenty-first century archival ecosystem. *Journal of Archival Organization*, 14(1-2), 13-34. https://doi.org/10.1080/15332748.2018.1503020

Woodbury, A. (n.d.) *What is an endangered language?* Linguistic Society of America. https://old.linguisticsociety.org/sites/default/files/Endangered_Languages_0.pdf

Zavalina, O.L., & Chelliah, S. L. (2022). Guest editorial: Digital language archives. *The Electronic Library*, 40(5), 505-506. https://doi.org/10.1108/el-11-2022-349

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

From pixels to perspectives: exploring perceptions of representation in character design

Nani Le Tang, Erin Eungyo Moon, Jose De Jesus Garrido Iniguez, Daniel Raul Mena Ramirez, and Daniel L Gardner

DOI: https://doi.org/10.47989/ir30iConf46939

Abstract

Introduction. This study presents findings about perceptions of how women characters are introduced in a sample of six best-selling games. Our analysis adds to ongoing conversations about representation in games and media, encouraging both designers and researchers to reflect on how perceptions and emotions related to representation shape interactions from the perspectives of gamers and non-gamers alike.

Method. We analysed 198 responses to a survey that asked participants questions about how they felt about short videos of character introductions and captured open-ended responses about the roles these characters occupied.

Analysis. We used descriptive and comparative statistics to identify basic trends in closed-ended survey responses amongst participants. We then identified, applied, and iterated on emergent descriptive qualitative codes designed around participant's open-ended responses.

Results. We identified a series of themes regarding cultural context, perceived characterization and character roles, and time and dialogue.

Conclusions. Our findings build on previous literature about character design and women's representation in games with analysis based on community perceptions and provide considerations for anyone designing characters and their introductions or working with diverse groups to make the medium and industry more inclusive.

Introduction

Between 65% and 70% of United States adults engage with at least one gaming platform (ESA, 2023; IGDA, 2022). The characters players take on or interact with in games are a key part of how people relate to and connect to the gaming medium. (Birk et al., 2016; Boellstorff, 2008; Freeman & Maloney, 2021; Huh & Williams, 2010; Klastrup & Tosca, 2009). Previous research examines sexism in the representation of these characters, in the character design process, and in communities that surround gaming, as well as the negative impact these phenomena can have on players (Beasley & Stanley, 2002; Behm-Morawitz & Mastro, 2009; Cullen & Ruberg, 2019; Downs & Smith, 2009; Leonard, 2003; Ruberg et al. 2019). In this paper, we begin to highlight the importance of studying how women characters are introduced in games.

While commonly associated with visual elements, '*character design*' is a multifaceted process describing the development of characters well-beyond their appearances, encompassing personalities, actions, and narratives. Though games are becoming more inclusive, many women characters are still designed with reductive gender roles. Scholars describe how women characters often present cliched tropes such as the '*damsel-in-distress*'—fragile, dependent, and awaiting rescue from their male counterparts—rather than as active heroic characters (Salter & Blodgett, 2017). Despite some recent changes in the Super Mario franchise, a popular long-running example is Princess Peach and her recurring pleas: '*help me, Mario!*', '*thank you, Mario!*', or '*Mario, you're my knight in shining armour!*' (Super Mario Wiki, n.d.). Recent games with women protagonists are not immune from relying on these tropes. For example, in a popular ongoing series, Final Fantasy XIII-2's main woman character is accompanied by two leading men upon whom she is highly reliant for protection and victory against enemies that she could not otherwise fight alone (Tompkins et al., 2020).

The reproduction of stereotypes and reductive gender roles can perpetuate harmful norms and perceptions of women and other groups. Shaw describes how representation and narratives influence what players even recognize as possible ways of '*being*' in the world (2017). Others identify how the hyper-sexualisation of female characters and reproduction of stereotypical gender roles can reflect the preferences of a gaming industry predominantly shaped and led by men (Breuer et al., 2015; Ochsner, 2017; Phan et al., 2012; Tompkins et al., 2020). Despite 46% of US adults that play games being women (ESA, 2023), less than 24% of industry workers are women (Zippia, 2022). These points describe a scenario where inherently sexist and stereotypical depictions and attitudes work recursively to shape the medium, industry, and communities while normalizing harmful consequences for women, and other marginalized communities.

Narratives structure the worlds characters live in and tell us more about who characters are beyond their appearance. Narratives enrich the gaming experience by conveying emotional, interpretive, and role-playing opportunities (Curtis, 2015; Lankoski, 2002). Qin et al. reiterate that '*without narrative, the existence of the game world and characters is meaningless*' (2009), as visuals without context cannot provide sufficient depth. Narratives can differentiate character or player experiences or encourage players to consider how in-game actions may affect how their characters grow, develop, or are perceived by others. Additional research is required to understand the impact of character narratives on players or potential players.

This paper analyses how 198 diverse survey respondents interpret the narrative introduction of women characters in a sample of best-selling games—both as protagonists and secondary characters. Aside from matters of project scale, we chose to focus on introductions to better understand the perceptions participants' have in their first impression/interaction with characters because first impressions can dramatically influence how someone is perceived. We aim to begin exploring the understudied relationship between women character narrative designs and how players perceive them. Our approach is guided by scholars who highlight the impact of reductive and harmful character designs on consumer perceptions and social norms (Cote, 2020;

Dill & Thill, 2007; Gestos et al., 2018; Lynch et al., 2016; Mou & Peng, 2009; Nakamura, 2019; Shaw, 2014; Tompkins & Martins, 2021), and those that try to incorporate diverse (and sometimes non-scholarly) perspectives in the analysis of games (Gardner & Corvite, 2023; Liu 2018; Reza et al., 2022). This study looks at the perceptions of diverse folks—along commonly considered demographic axes and how frequently they report playing games. Our findings provide several considerations for anyone designing characters and their introductions or working with diverse groups to make the medium and industry more inclusive.

Related work

Based on an initial review of 108 articles across disciplines related to women's representation, character design, and sexism in gaming. These articles are organized into three themes: sexism in character design, sexist norms in gaming communities, and women's representation and social perspectives.

We identify a gap in the literature; while there are many *scholarly* content analyses and interpretations of characters narratives, there is a dearth of research centring how diverse *players or potential players* perceive and interpret these woman character narratives.

Sexism in character design

Many scholars discuss the sexist ways women are depicted in and around video games (Beasley & Standley, 2002; Behm-Morawitz & Mastro, 2009; Carvalho & Cappelli, 2018; Cote, 2018; McArthur, 2018; Mou & Peng, 2009). This includes quantitative and qualitative explorations of biased, stereotypical, reductive, and potentially harmful visual and narrative depictions of women in games.

Many scholars examine sexist depictions by focusing on the frequency of women characters and the qualities of their visual appearance within videogames and associated paratexts (Burgess et al., 2007; Dill & Thill, 2007; Gardner & Tanenbaum, 2018; Gestos et al., 2018; Miller & Summers, 2014; Shaw & Friesem, 2016). Although scholars observe an increase in the amount of women characters present in videogames (Lynch et al., 2016), men are still much more frequently main characters compared to women when present (Gardner & Tanenbaum, 2018; Maclean, 2016). Other scholars observe how women are still frequently depicted with unrealistic bodies and exaggerated sexualization (Downs & Smith, 2010; Glaubke et al., 2001; Jansz & Martis, 2007; Tompkins & Martins, 2021), complicating the quality of increased representation.

Scholars also examine narratives, roles, and dialogues of women in video games (Butt & Dunne, 2017; Hart, 2017; Gray et al., 2017; Kennedy, 2002; MacCallum-Stewart, 2014; Masso, 2011; Perreault et al., 2018; Santonicollo et al., 2023; Sherman, 1997; Ugolotti et al., 2020). Many of these studies suggest women characters often give insignificant or demeaning roles in relation to main (men) characters (Friedberg, 2015; Masso, 2011). For example, Heritage describes how men in the popular Witcher series more often have active roles with more power (e.g., kings or military leaders) while women more often serve as advisors and onlookers (2022).

Despite these negative trends, several scholars also examine more contemporary characters who challenge stereotypical depictions (Allen, 2023; Lankoski, 2004; Lucas, 2019; Solska, 2022). For example, Solska identifies a sample of three games that present women characters in narratively focal roles in which they interact with other characters and environments instead of only being submissive and passive (2019). These studies encourage us to consider what and how other aspects of characters may be sexist, and how to begin to confront the issue beyond characters' surface appearance.

Sexist norms in gaming communities

A wide variety of scholars identify how norms can dramatically shape how people experience social, professional, and recreational communities such as those that surround games (Auerbach, 2014; Campbell, 2014; Cote, 2017; Edidin, 2014; ESA, 2023; Fox & Tang, 2014; Gonzalez et al, 2014; Gray et al., 2017; Hart, 2017; Koh & Das, 2017; Lenhart et al., 2008; Leonhardt, 2021; Mou & Peng, 2009; Ochsner, 2017; Reza et al, 2022; Shliakhovchuk & Muñoz García, 2020; Tompkins et al., 2020; Weber, 2012; Williams, 2003), and can contribute to the stereotypes discussed in the previous section. We organize our research with this theme that describes norms, attitudes, and structural forces that deter women from accessing the digital game medium, its production, and associated communities to the fullest.

Several studies suggest the predominantly male game industry perpetuates a sexist workplace culture as well as negative or harmful portrayals of women (Gray et al. 2017; Mou & Peng, 2008; Tompkins & Martins, 2021. Gaming has too often been considered '*an activity created by men and for men*' (Carvalho & Cappelli, 2018), with many women professionals testifying to workplace discrimination and/or harassment (Weber, 2012). Sexism in the gaming industry makes it harder for women to succeed professionally and facilitates '*the sexualization of female game characters*' (Tompkins et al., 2020). These studies help support the persistent claim that increasing equity and inclusion, and hiring more women, are necessary in the design process to address the sexistnormative industry environments that perpetuate sexist representations (Campbell, 2018; Doerrer & Campain Live, n.d.; Elderkin, 2023; Kennedy, 2018; 2002; Passmore et al., 2017; Qin et al., 2009; Reza et al., 2022; Solska, 2022; Wood & Wood, 2008).

Perceived market factors can drive sexist representations of women in games and reproduce past narratives as a risk management strategy (Stermer & Burkley, 2015; Ugolotti et al., 2020). Tompkins and Martins describe demands for a '*success formula*' (e.g., Figure 1) to guarantee return of investments based on repeating past (men-dominant) successes (2021). However, this approach heavily depends on assumptions about the '*real world*' that do not always align with the actual population of players (Duggan, 2015a; 2015b; Lenhart et al., 2008; Lenhart et al., 2015; Passmore et al., 2017; Shaw, 2009; 2014).

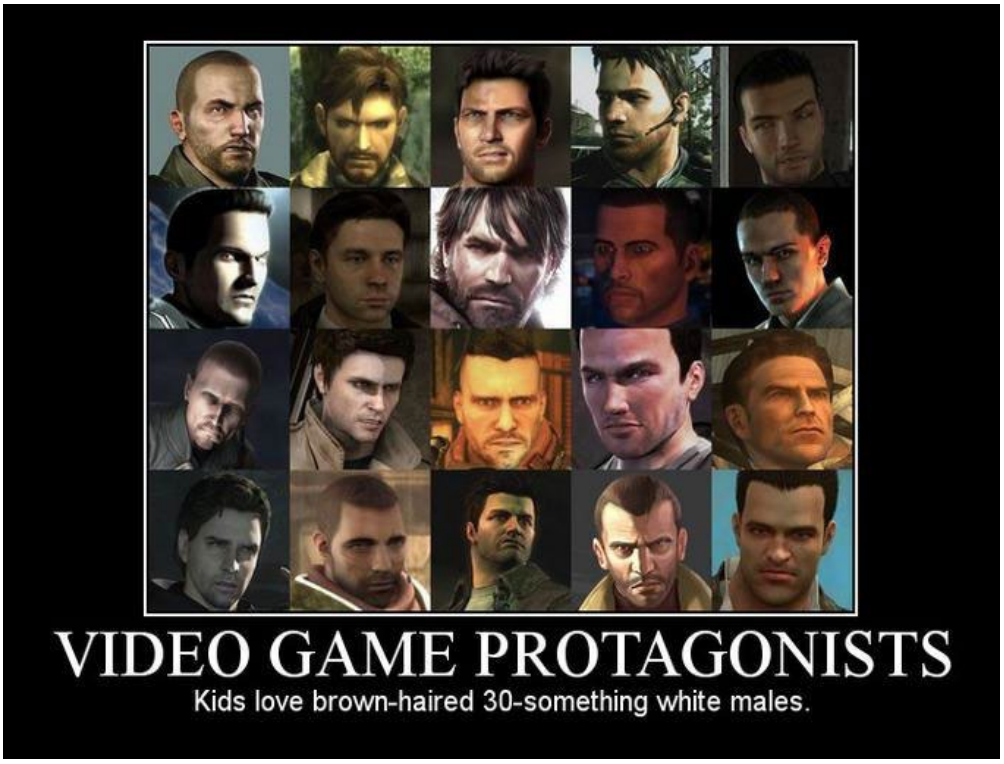


Figure 1. Example of a common meme characterizing dominance of white men main characters (n.d.)

Player-based gaming communities are also historically discriminatory towards women. There is a common refrain that '*girls cannot play, are mechanically unskilled, and owe their achievements to a male counterpart who has helped them get there*' (Koh & Das, 2017). Scholars identify how women players are regularly subject to harassment and sexist environments (Aghazadeh et al., 2018; Bezio, 2018, Campbell, 2014: Condis, 2015; Edidin, 2014; Massanari, 2017; Mortensen, 2018; O'Leary, 2012; Wingfield, 2014). with some suggesting lax consequences and scrutiny of such norms contributing to their proliferation (Fletcher, 2012). Fox and Tang identify how endorsing broader masculine norms is a predictor of sexist attitudes about women playing video games (2014). While Cote reveals that women players are often capable of managing online game-related harassments, it is an inequitable burden that they need to adopt coping strategies at all simply to participate in the medium (2017).

Women's representation and social perspectives

Despite the wealth of studies identifying poor representation in games described in section 2.1 and beyond, there are surprisingly few studies examining how players and people surrounding the medium themselves perceive and interpret representation. Beyond the general influence on formal and informal norms discussed in section 2.2, negative visual and narrative portrayals can be linked to harmful outlooks on life (Stermer & Burkley, 2015). We organize studies that analyse the influence of the representation on people and community perceptions (emphasizing harmful effects) or that analyse how people or communities themselves perceive or interpret that representation.

Studies show how representation in video games can influence player attitudes and perceptions about gender (Bowey, 2017; Liu, 2018; Mou & Peng, 2009; Shaw, 2014; Stermer & Burkley, 2015; Yao et al., 2010). Frequent engagement with sexist games links to sexist attitudes in players and negatively impacts women's well-being beyond games (Fox & Bailenson, 2009; Gabbiadini et al., 2016; Scharrer, 2004; Summers & Miller, 2014). In a meta-analysis, Gestos et al. demonstrates how

increases in objectification in games can diminish women players' self-efficacy and well-being while promoting sexist views in male players (2018).

Though fewer in number, some studies examine how players or potential players themselves perceive, interpret, and respond to representation in games (Desai et al., 2017; Stermer & Burkley, 2015; Gardner & Hacker, 2022; Tomkinson et al., 2020). Kondrat found that 76.9% of participants in their study recognize negative stereotypes in women characters in games, generally, and 31.62% identify sexual objectification specifically, with a majority of participants desiring less stereotypical characters (2015). Other scholars identify how the interpretations of character representation, and their influence can vary between different player demographic identities (Gardner & Corvite, 2023; Liu, 2018; Nguyen et al., 2020; Reza et al., 2022). Our study further explores community perceptions focusing on women character narratives rather than focusing just on researcher-centred content analysis of physical appearances.

Methods

We developed a survey following Chapman University IRB approval, containing a combination of closed and open-ended questions. Using a sample of games, we focused on understanding how racially and gender diverse participants perceived and interpreted the narrative introduction of women characters, how those characters might resemble real life women, as well as participants' broader feelings about the characters. We analysed our survey results quantitatively and qualitatively using descriptive and comparative statistics, and emergent descriptive codes, to identify basic trends in overall responses and between participant groups.

Our approach was organized around a primary overarching research question:

RQ1. How do diverse community members' interpretations of how women videogame characters' introductions reflect/express broader social norms around gender?

Addressing every personal and cultural dimension of this question is beyond the scope of this study. However, our study begins to address this question by examining how diverse participants in terms of race, gender, and frequency of play—likely from North America—perceive and interpret six women's introductions of best-selling narrative games.

Survey design

Our primary analysis is of participant impressions rather than the videos themselves. That is, rather than us defining or assigning sufficiency to—for instance—introduction duration and asking participants how strongly they agreed with us, our key data are participants' own interpretations of sufficient understanding, within the scope of provided clips. We chose a survey during our preliminary stages to reach larger and more diverse audiences and to better understand the state of the medium. Our survey was organized into three sections on a Google Form. The first section asked participants about demographic characteristics including age, primary language, gender, race, gaming habits and aspects they felt make characters more relatable. The second section showed participants one-minute video game clips introducing six sampled characters outlined below in section 3.2. Following each clip, we asked participants to respond to four statements about their impressions of each character/introduction based on time, dialogue, how strongly they felt the characters resemble any women they know in their life, and their perception of the character's importance to the game's narrative. These questions utilized Likert scales to evaluate how strongly participants agreed or disagreed with statements about our sample of characters and their introductions. The third section asked questions about participant thoughts on the roles of the sampled characters and any other remaining general thoughts or feelings.

Game/clip selection

In November 2023, we selected the three best-selling '*story rich*', single-player games with women protagonists and the three best-selling games with men protagonists, where women characters were introduced (with dialogue) within the first two hours of gameplay. We sampled from Steam because it is the largest digital games retailer, accounting for 61% of all full game revenue from new releases in 2023 (Fragen, 2024). We did this by going to their digital store and selecting '*top sellers*', filtering by the tags above, and identifying the best-selling games by units sold on Steam. We chose to look at recent mainstream games from 2023 to capture perceptions about a relatively current snapshot of the state of the medium. We narrowed down our game selection to games meta-tagged with story-rich and single-player as these generally suggest heavier character narratives and/or story-driven goals. We excluded games that had deuteragonists or customizable main characters, non-human characters, and games without clear dialogue for women characters in the first two hours of the game. We made these exclusions to maintain a clear focus on the representation and narrative introduction of women characters.

We chose to focus on one-minute clips to be respectful of participants' time by providing shorter length videos, only capturing first scenes that introduce the characters. For games with women protagonists, we found clips of their introductions/first dialogues using commentary-free gameplay walkthroughs, or '*let's play*', on YouTube. Similarly, we also identified clips for the first appearances/introductions of supporting women characters with dialogue within the first two hours by first watching walkthroughs to confirm on YouTube.

We wanted to observe potential differences in perceptions between games led by women protagonists and games with women but led by men protagonists. We chose sub-samples of three of each game-type and six games total to reach a minimum number of examples to make comparisons while also trying to be considerate of participants' time and the length of the survey.

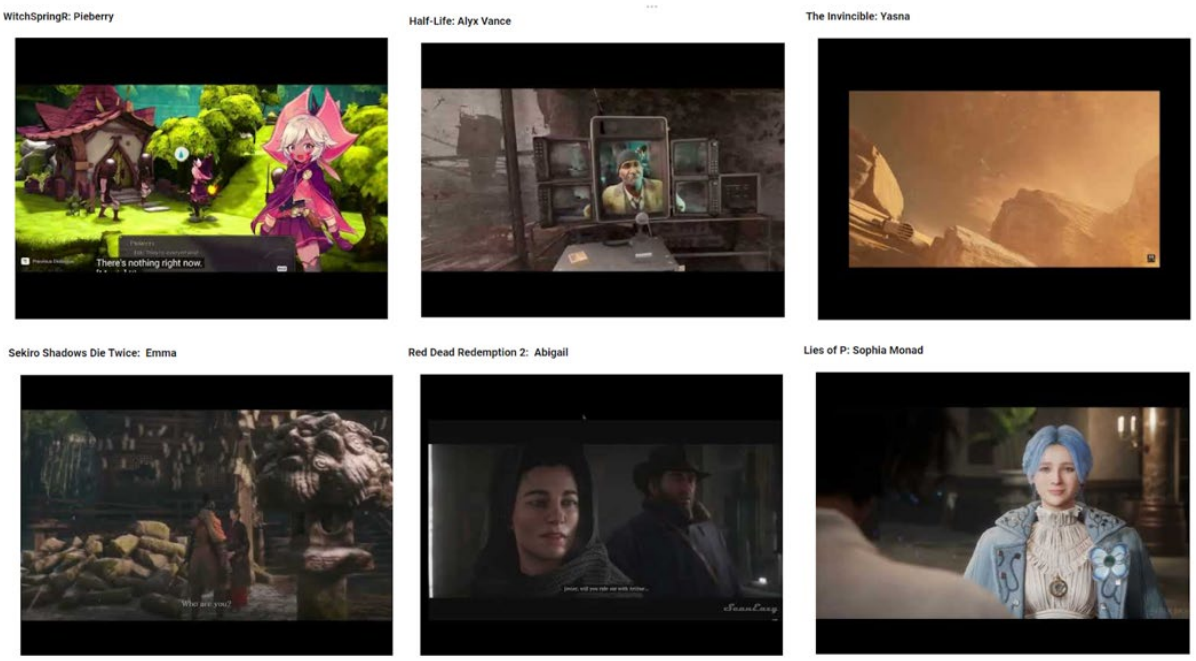


Figure 2. Illustrates the six gaming clips, women-led (top) and men-led (bottom), included in the survey

Recruitment

To be eligible, the only requirement was for participants to be at least 18 years old and speak English—because the survey was in English—regardless of their gaming experience. However, we were especially interested in the perspectives of women and anyone who feels historically underrepresented in media.

Our team digitally and physically distributed the survey in three waves. The first two waves involved social media platforms such as Discord, email lists, Facebook, Instagram, LinkedIn, and Slack. Our team members each represent unique interdisciplinary and demographic backgrounds and—at time of data collection—they were based at four different institutions in California, Indiana, and Texas. We began with personal networks and moved on to forums and community groups related to gaming—either public groups online or associated with student groups on our respective campuses. To further increase participant diversity, we made additional efforts to include communities centred on women and gaming such as the '*women in gaming*' discord or inclusive gaming network and the game HERs. The third wave involved posting flyers and recruiting faculty at our respective institutions to physically solicit additional possible participants. Because our sampling relied heavily on university groups centred in North America, participation was biased toward college-based/aged communities, likely residing in North America.

Data analysis

We began by using descriptive and comparative statistics to identify basic trends in closed-ended responses by all participants and between participant groups (e.g., race, gender, or frequency of play). We then identified, applied, and iterated on emergent descriptive qualitative codes for open-ended responses.

We ran three waves of Welch's t-tests focused on responses to the second section of the survey. Responses to the Likert scales were converted to integers 1 to 5 to calculate averages and make comparisons, with 1 representing strongly disagree, 5 representing strongly agree, and 3 representing a neutral response.

The first wave consisted of high-level t-tests within and between groups based on identified race and gender. Specifically, we compared responses between participants who identified as White and those who did not, and between those who identified as men and those who did not. We chose to start with these separations to see if we could observe any differences in perception between those who are historically well-represented in games and media and those who are not. That is, white men are often more represented than other folks in games (Gardner & Tanenbaum, 2018; Passmore et al., 2017; Williams et al., 2009), and in the game industry (AGA, 2023). We calculated each participant's mean response to all six copies of the repeating questions in the second section of the survey and tested them between groups to identify any significant differences. We compared responses by these groups across all six games, across men-led games, across women-led games, and for each individual game. One participant who preferred not to state their gender was only included in analysis based on identified race and of total participants.

The second wave involved the same comparisons across games and sub-samples of games, but tested responses based on intersectional demographic identifiers combining two variables such as race, gender, and frequency of stated gaming activity (e.g., every day, a few times a week, a few times a month, rarely, never). We did, however, still primarily with a focus of comparing the '*default*' against everyone else. For example, we compared responses between white-men and not white men, white-women, and not white women, and those who identified as white and as playing a particular frequency and those who did not identify as white who played that frequency.

We conducted the third wave after qualitative review of open-ended responses, specifically participant comments relating to the cultural context of characters (discussed more below). In this

final round of tests, we compared responses to the introductions in Half-Life: Alyx, Red Dead Redemption 2, and The Invincible, realizing based on responses that they were Western (North America and Europe) developed games, and Asian-developed games: Lies of P, Sekiro, and Witch Spring R.

We created and applied emergent, descriptive codes to responses to the open-ended questions in the final section (Saldaña, 2015). Each author independently read and coded open-ended responses to independently identify themes. The authors then came together to discuss, aggregate, and iterate on thematic codes, and consider how these themes and participant quotes related to quantitative data.

Findings

We organized our findings around comparisons between responses of singular or intersectional demographics of identified gender and race and a series of related qualitative themes we found in our open-ended responses. Our three key findings describe participants' perceptions regarding cultural context, characters' roles, and the presentation of character introductions based on time and dialogue.

Survey responses came from a more diverse than representative sample. The most represented gender category participants most-identified as was Woman (49.2%), followed by Man (41.6%), and Nonbinary (7.1%). Table 1 shows a detailed breakdown of participant gender demographics for our study. The most represented racial category participants identified as was White (32.8%), followed by Latine/o/a (19.7%) and Mixed/Multi-racial (14.7%) (Table 2). Although we asked about age, primary language, and frequency of play, we did not identify any key findings associated with these variables.

Gender	n (%)
n	198 (100%)
Women	97 (48.99%)
Men	83 (41.92%)
Nonbinary	17 (8.59%)
Prefer not to identify	1 (0.51%)

Table 1. Gender demographics of survey participants displayed from greatest to least number of participants (response to Q2 in survey)

Racial/Ethnic Demographic	n (%)
n	198 (100%)
White	65 (32.8%)
Latine/o/a	39 (19.7%)
Mixed	29 (14.7%)
East Asian	21 (10.6%)
South-East Asian	21 (10.6%)
Black/African American	15 (7.6%)
South Asian	7 (3.5%)

Table 2. Racial demographics of survey participants displayed in order of prevalence

Cultural Context

Our quantitative results showed that all participants, on average, were more likely to know at least one woman similar to the characters that appeared in games developed in a Western context compared to those developed in an Asian context. Relatedly, we observed multiple participants not only acknowledging the role of a developer's culture in open-ended responses, but also identifying tropes and stereotypes that may rely on cultural reference such as one participant's use of '*kawaii*', a Japanese trope for '*lovable*' or '*cute*'.

When asked how strongly they agreed or disagreed with whether characters were similar to at least one woman they knew in real life, participants, on average, were more likely to agree for characters developed in a Western context compared to those developed in an Asian context ($p=0.0190$). Technically, participants neither agreed nor disagreed, on average, that they knew at least one woman similar to the characters created by Western developers ($\bar{x}=3.16$), However, they disagreed, on average, with knowing at least one woman similar to those in games created by Asian developers ($\bar{x}=2.24$).

We also observed a variety of open-ended responses that directly recognized cultural components of video game developers as well as their impact on woman character development. For example, three participants explicitly mentioned Japanese culture and context for development/consumption when describing Witch Spring R ('*Japanese-made*'; '*J-game for a J-Audience*'; '*Her creators might have hailed from Japan*'). Another described this character saying she is '*cute and her speech being kawaii.*'

We use cultural context to describe the customs, norms, values, and language surrounding a community. Our findings indicate participants implicitly and explicitly reacted to cultural influences in the sampled games, notably, when they live in the same cultural context as the game's developers. For developers, these results should indicate how their experiences may influence content, and how it may resonate with different players.

Perceived characterization and character roles

We observed that participants—especially in aggregate—often explained their perceptions of characterizations and character roles in sampled introductions using oppositional descriptors. We coded 62 comments out of 110 open-ended responses (56.36%) with one or more of the following codes: Strong vs Weak, Main vs Support, and Important vs Basic. Strong vs Weak Characterization and Main vs Support were our two most common codes applied overall. Moreover, participants used descriptors from multiple oppositional axes, perhaps describing a character as weakly characterized, in a supportive role, and basic.

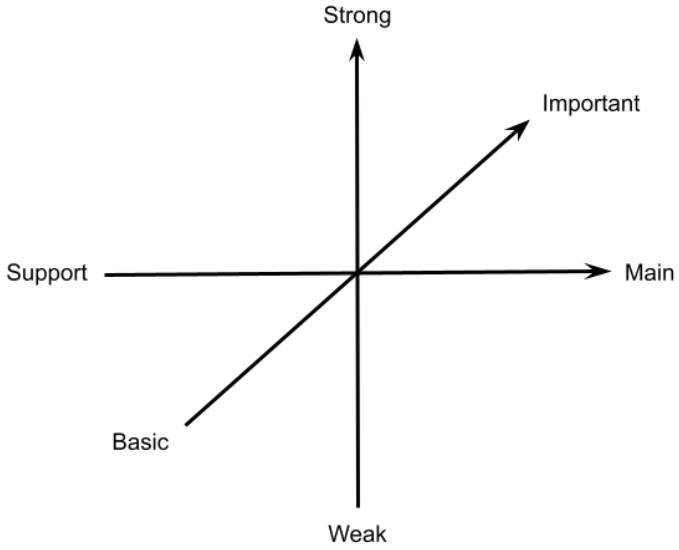


Figure 3. Three-dimensional axes of oppositional descriptors

We used strong vs. weak characterization to describe participant comments about the general depth and complexity of game characters. Participants did not always use the words '*strong*' or '*weak*', but referring in some way to general complexity, detail, or lack of either were cause for applying the code (e.g., '*very strong characters and personalities*', '*headstrong*', '*[not for use of strong, but strong impression of character]*', '*pretty vague*', '*very little detail*', '*lack depth*', '*understated or flat*'). Both similarly and conversely, we applied Main vs Support to participants using explicitly contextual or game-mechanical terms to describe characters' relative roles (e.g., main character, support character; '*helpers or one-off*', '*assistive to the main character*'). We applied Important vs. Basic to participant comments discussing a character's perceived overall relevance to the narrative (e.g., primary cast member or once-met NPC; "helpers or one-off" again).

We should also note that while we saw comments on both sides of these oppositions, participants more frequently described characters in our sample as weak, supportive, and/or basic. For example, from the 38 responses coded with '*strong vs weak characterization*', 32 contained comments about weaker characterizations while only 11 contained comments about stronger characterizations; we had five responses that discussed both ends at the same time. We received multiple similarly worded comments relating women characters to the more negative sides of oppositions and men characters to the more positive sides:

> *It seems like for most of the female characters presented here serve a role to aid the protagonist/player in some way, making them important to plot, but don't serve any further purpose. Then there are the female characters that we actually play as and see their perspective. They seem to exert more personality to a degree.*'; '*It seems like most female characters were side characters that help the main protagonist. Though some*

clips showed strong female leads which I believe to be was nice to see'; '*They all seem to be like not thoroughly represented to me.*'; '*Most of the time seemed to be side characters with little depth beyond what is required to give a mission*'; '*Apart from a few exceptions most of the presented roles seem to be assistive to the main character.*'

Although participants occasionally mentioned some '*strong*' women characters, an overwhelming amount of responses discussed that all or most women characters portrayed in these games seemed to have shallow characterization and assistive roles.

Time and dialogue

Overall, we observed that all participants, on average, felt neither the time nor dialogue of the sampled introductions were sufficient or insufficient enough to understand sampled characters. In fact, we saw near consensus across all six games—with several p-values above .95—with participants neither agreeing nor disagreeing (averaging 3.11 and 3.24, for time and dialogue respectively) that time or dialogue of the introductions were sufficient to even begin to understand these characters. We did not find a significant difference between responses to these questions based on demographics, individual games, across all six games, or whether games were women-led or not. Although neutrality itself is not a finding, in this case it presents an inconclusive result when attempting to analyse participant perceptions.

Despite the neutrality displayed, on average, in our closed-ended questions, many open-ended responses discuss the seemingly incompleteness and length of introductions or the depth and quality of the dialogue. Several participant comments coded with time—to refer to explicit references to the concept—expressed that the introductions felt '*understandably short*', '*to the point*', '*rushed*', '*unexplained*', *or* '*did not have much introduction time.*' Similarly, many comments coded with dialogue—for the same reason as time—described it as '*pretty vague*' *or* clearly stated that '*none [of the characters] had sufficient enough dialogue.*' The inconclusive contrast between expressive open-ended responses and neutral quantitative responses suggests there may have either/both been underlying issues with our survey construction, or the content of the clips we used themselves, discussed below

Limitations and opportunities

We identified two main limitations during data collection and analysis. These included a non-generalizable level of diversity in our participants and details of sampling.

Partly due to our recruiting method, the majority of respondents were 18-25 years old, likely from North America, and identified English as their first language, meaning our sample is not representative. However, the sample is more diverse along racial and gendered lines than a representative sample, meaning we were able to observe the perceptions of historically underrepresented groups at a higher rate. Still, there is an opportunity for future studies to sample wider age groups and broader cultural and geographic communities.

Despite challenges with clip duration and selection, our selection of only six games ranked as best-sellers on Steam November in 2023 does limit analysis and leaves room for potential biases. Our sample, and analysis represents a snapshot of the medium and market. Greater sampling, or repeat studies are necessary for broader or more historical analysis of the medium and market over time. A future project could focus on people in a different region or context (e.g., Asia, Europe, or S. America) or experiment with other modes of delivering clips with the same study to see if the same findings are replicable.

Discussion

Cultural context

Our findings reinforce the notion that the perceptions and assumptions of developers are embedded in the games they make—and that players and non-players alike are often aware of this influence, if only implicitly. We did not in any way indicate the development context of games in our survey, nor did we intentionally sample three western-developed games and three Asian-developed games. Yet, participants observed differences in the characters and narratives between these groups of games and found they were more familiar with women characters developed in their own likely cultural context. These findings highlight how important it is to consider cultural contexts when developing or analysing character designs and narratives—especially when they are introduced or creating their first impressions—or when examining how potential players perceive those characters.

Our study suggests that if game makers want to develop characters that resemble women from more diverse cultural contexts, they need to include more diverse perspectives in the development process–specifically in roles with significant influence on the creation of character narratives or their narratives.

The use or production of culturally-linked stereotypes and tropes may also influence how potential players—and those around them—perceive game characters. Our analysis highlights how sexist representations previously analysed by many scholars described in sections 2.1 or 2.3, such as Summer and Miller's emphasis on oversexualization of women in game-related media compared to men, may be more closely tied to regional, cultural contexts than previous scholars have acknowledged (2014). Better understanding how developer and player perceptions influence—or are impacted by—characters is important for mitigating the perpetuation of stereotypical conceptualizations that reinforce negative representations of certain groups in games.

Perceived characterization and character roles

Our three-axes visualization allows future researchers to pivot from Likert scales to ask different, more specific, sorts of questions as part of surveys, interviews, or focus groups. How participants, players, or potential players might *chart* their perceptions allows us to ask more precise questions about, for example, '*how*' and '*why*' they may position certain characters, or how/why they position them relative to each other—potentially inviting deeper, multi-dimensional insights even in relatively brief interactions.

Such frequent use of specific oppositional descriptors also urges researchers to question *why* we are seeing this trend in the first place. Perhaps participants genuinely perceive men and women as oppositional descriptors, or perhaps women characters are developed and presented to fit within the three axes. Either way, our visualization can help encourage research questions that reflect on these oppositions, their origins and beg a deeper understanding of game consumer's perception.

Participants leaning toward describing presented portrayals of women characters on the weaker, supporting, basic ends of identified axes suggests that game developers should be mindful of their perceptions on gender when developing characters. As discussed in the previous section, it is likely that perceptions of game developers are embedded in their creations. Hence, bringing more women's perspectives to the workspace may help towards robust character development and overall success of the game if game developers aim to improve women representation in video games.

Time and dialogue

The inconclusiveness we observed on the '*sufficiency*' of time and dialogue in introductions may stem from the durations of clips, the content of clips, or how we described/presented/asked about clips/introductions. Because we were presenting participants with 6 clips of footage, we

attempted to make clip durations digestible. It is tempting to consider adjusting clip durations—perhaps making them longer—as an immediate fix. Our focus on clip quantity may have influenced the neutral responses, on average, perhaps due to participants hurrying through or due to clips actually being too short but participants being uncertain how to respond. However, because participants perceived them as neither sufficient nor insufficient, simply changing or increasing the duration of clips could remain ineffective in identifying participant perceptions. Alternatively, the nature of the in-game introductions our clips contain may be the issue. In this case, participants' neutral consensus could be communicating, on average, that time and dialogue simply did not positively or negatively affect their understanding of the characters. The inconclusiveness, then, may be driven by the quality of dialogue, narrative, or overall characterization. As an additional note, we also must acknowledge in hindsight that participants may have understood the use of '*introduction*' in two ways: the character's in-game treatment or the clip itself embedded in the survey.

Modifying question structure and asking additional questions to clarify how time and dialogue affect participant perceptions and better acknowledging how the quality of character introductions can affect initial understandings of the character can help similar studies avoid similar inconclusive results. Rephrasing questions can reduce potential misinterpretations of '*introduction*'. Asking high-level questions about how participants consider the role of time and dialogue to influence their experiences, or about the quality of sampled introductions can provide better understanding of how they perceive those introductions. However, it bears mention that the issue may simply be issues with the inherent quality of those introductions themselves and a need for developers to produce introductions that allow participants to develop initial understandings of characters more effectively.

Conclusion

This study contributes to a better understanding of how potential players perceive and feel about representation in games and media, using a sample of six introductions of women characters from best-selling games. Our analysis highlights three key findings describing how participants' perceptions reflect and respond to cultural context, reflect multidimensional understandings of character roles, and are potentially influenced by the limitations of time and dialogue. The first two findings provide considerations for future researchers and game developers aiming to develop complex and culturally diverse women characters. Our findings suggest that if game developers want players from more diverse cultures to connect to their game, they need more intercultural design leadership, and consider localization as *part* of design not a translation of it. The last finding helps researchers to structure similar studies to observe perceptions quantitatively more precisely and consider how the quality of character introductions may impact understanding.

Narratives can tell us more about who characters are beyond their appearance–especially the introduction where first impressions are made by the player. These introductions may set their understanding of a character's potential role, personality, story, or their perceived importance to the game. Introductory dialogue influences perceptions of how these women are presented, and the state of the medium inhabited by these characters. Scholars such as Shaw have described how representation and narratives influence what players even recognize as possible ways of '*being*' in the world (2017). Attending to the themes we identify can help anyone designing or researching characters and their introductions or working with diverse groups to make the medium and industry more inclusive.

Acknowledgments

The iSchool Inclusion Institute, Kayla Booth and Joe Sanchez who connected us all, and supported our team from start to end of this project.

About the authors

Nani Le Tang is an undergraduate student studying Informatics with a focus in UX Design and Design Strategies at The University of Texas at Austin. Her research interests are in understanding and improving the experiences of people in media and technology, focusing on user research, interfaces, and accessibility. She can be contacted at nanitangux@gmail.com

Erin Eungyo Moon is a graduate student in Heinz College, Carnegie Mellon University, studying Information Security Policy and Management. They received their B.S. in Social Justice Informatics in the University of Texas at Austin. They can be contacted at emoon@andrew.cmu.edu

José De Jesús Garrido Iñiguez is an undergraduate student specializing in Game Design and Development at Indiana University Indianapolis. He primarily focuses on producing creature or character concept art during preproduction and furthering designs into the implementation stages. They can be contacted at jgarrid@iu.edu

Daniel Raul Mena Ramirez is a Technology Support Specialist at eMetric. He received his bachelor's degree in Information Systems and Cyber Security from The University of Texas at San Antonio. His research interests are in the documentation of the influence technology has on social expectations. He can be contacted at dan.ram.3611@gmail.com

Daniel L Gardner is a lecturer of computing and game design at Robert Gordon University. He examines inequitable interactions enabled or enforced by interfaces and infrastructures on the periphery of digital games and media. He primarily examines transactions of authority and exclusion mediated by access controls, character configuration, and physical interfaces/controllers that disproportionately affect historically marginalized communities.

References

Aghazadeh, S.A., Burns, A., Chu, J., Feigenblatt, H., Laribee, E., Maynard, L., Meyers, A.L., O'Brien, J.L. and Rufus, L. (2018). GamerGate: A case study in online harassment. Online harassment, 179-207.

Allen, M.L. (2023). "How is it that Someone Like You: paragon, Damn Near Saint; Could Love this World so Damn Much but No One in It?": Horizon Zero Dawn's Aloy and the Disruption of Hetero-Amatonormative Benevolent Sexism in Female-led Games. Games and Culture.

American Gaming Association (AGA). (2024). 2024 Gaming Industry Workforce Diversity Report.

Auerbach, D. (2014). Letter to a young male gamer. Slate Magazine.

Beasley, B. and Collins Standley, T. (2002). Shirts vs. skins: Clothing as an indicator of gender role stereotyping in video games. Mass Communication & Society, 5(3), 279-293.

Behm-Morawitz, E. and Mastro, D. (2009). The effects of the sexualization of female video game characters on gender stereotyping and female self-concept. Sex roles, 61, 808-823.

Bezio, K.M. (2018). Ctrl-Alt-Del: GamerGate as a precursor to the rise of the alt-right. Leadership, 14(5), 556-566.

Birk, M.V., Atkins, C., Bowey, J.T. and Mandryk, R.L.May. (2016). Fostering intrinsic motivation through avatar identification in digital games. In: Proceedings of the 2016 CHI conference on human factors in computing systems, 2982-2995.

Boellstorff, T. (2008). Coming of age in Second Life: An anthropologist explores the virtually human. Princeton University Press.

Bowey, J.T., Depping, A.E. and Mandryk, R.L. (2017). May. Don't talk dirty to me: How sexist beliefs affect experience in sexist games. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 1530-1543.

Breuer, J., Kowert, R., Festl, R. and Quandt, T. (2015). Sexist games= sexist gamers? A longitudinal study on the relationship between video game use and sexist attitudes. Cyberpsychology, Behavior, and Social Networking, 18(4), 197-202.

Burgess, M.C., Stermer, S.P. and Burgess, S.R. (2007). Sex, lies, and video games: The portrayal of male and female characters on video game covers. Sex roles, 57, 419-433.

Butt, M.A.R. and Dunne, D. (2017). Rebel girls and consequence in life is strange and the walking dead. Games and Culture, 14(4), 430-449.

Campbell, C. (2014). Sarkeesian driven out of home by online abuse and death threats. Polygon.

Campbell, C. (2018). Game companies are failing on diversity, according to new report. Polygon.

Carvalho, L.P. and Cappelli, C. (2018). Sexism and League of Legends: NFR aesthetic analyses. In: Anais da V Escola Regional de Sistemas de Informação do Rio de Janeiro, 38-45.

Condis, M.A. (2015). The politics of gamers: identity and masculinity in the age of digital media. University of Illinois at Urbana-Champaign.

Cote, A.C. (2017). “I can defend myself” women’s strategies for coping with harassment while gaming online. Games and culture, 12(2), 136-155.

Cote, A.C. (2018). Writing “Gamers” the gendered construction of gamer identity in Nintendo power (1994–1999). Games and Culture, 13(5), 479-503.

Cote, A.C. (2020). Gaming sexism. In Gaming Sexism. New York University Press.

Cullen, A.L. and Ruberg, B. (2019). August. Necklines and'naughty bits' constructing and regulating bodies in live streaming community guidelines. In Proceedings of the 14th International Conference on the Foundations of Digital Games, 1-8.

Curtis, E.D. (2015). In the game of patriarchy: The damsel in distress narrative in video games. The University of Alabama.

Desai, N., Zhao, R. and Szafron, D. (2017). Effects of gender on perception and interpretation of video game character behavior and emotion. IEEE Transactions on Computational Intelligence and AI in Games, 9(4), 333-341.

Dill, K.E. and Thill, K.P. (2007). Video Game Characters and the Socialization of Gender Roles: Young People’s Perceptions Mirror Sexist Media Depictions. Sex Roles 57, 851–864.

Doerrer, B. and Campaign Live. To reach diverse audiences, game developers must transform from within" from Campaignlive.com, https://www.campaignlive.com/article/reach-diverse-audiences-game-developers-transform-within/1804774.

Downs, E. and Smith, S.L. (2010). Keeping abreast of hypersexuality: A video game charTo reach diverse audiences, game developers must transform from within | Campaign USacter content analysis. Sex roles, 62, 721-733.

Duggan, M. (2015). Gaming and Gamers. PEW Internet Research.

Duggan, M. (2015). Which Americans play video games and who identifies as a 'gamer.' PEW Internet Research.

Edidin, R. (2014). She was harassed by a games reporter. Now she's speaking out. Kotaku.

Elderkin, B. (2023). GDC 2023 State of The Game Industry: Devs Weigh in On the Metaverse, Player Toxicity, And More" GDC conference.

Electronic Software Association (ESA). 2023). 2023 Essential Facts about the Video Game Industry.

Fletcher, J. (2012). Sexual harassment in the world of video gaming. BBC News Magazine.

Fox, J. and Bailenson, J.N. (2009). Virtual virgins and vamps: The effects of exposure to female characters' sexualized appearance and gaze in an immersive virtual environment. Sex roles, 61, 147-157.

Fox, J., and Tang, W.Y. (2014). Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. Computers in human behavior, 33, 314-320.

Fragen, J. (2023). Steam's top 10 games captured over 60% of revenue in 2023, https://venturebeat.com/games/video-game-insights-steam-pc-revenue-consolidation/.

Freeman, G. and Maloney, D. (2021). Body, avatar, and me: The presentation and perception of self in social virtual reality. Proceedings of the ACM on human-computer interaction, 4(CSCW3), 1-27.

Friedberg, J. (2015). Gender games: A content analysis of gender portrayals in modern, narrative video games.

Gabbiadini, A., Riva, P., Andrighetto, L., Volpato, C. and Bushman, B.J.: Acting like a tough guy. (2016). Violent-sexist video games, identification with game characters, masculine beliefs, & empathy for female violence victims. PLoS one, 11(4).

Gardner, D.L. and Corvite, S. (2023). First Impressions: Effects of Representation on Video Game Covers.

Gardner, D.L. and Hacker, K. (2022). Prosthetic Metaphors, Rejection, and Representation in Games. In Proceedings of Digital Games Research Association Annual Meeting.

Gardner, D.L. and Tanenbaum, T.J. (2018). Dynamic demographics: lessons from a large-scale census of performative possibilities in games. In Proceedings of the 2018 CHI Conference on human factors in computing systems, 1-12.

Gestos, M., Smith-Merry, J. and Campbell, A. (2018). Representation of women in video games: A systematic review of literature in consideration of adult female wellbeing. Cyberpsychology, Behavior, and Social Networking, 21(9), 535-541.

Glaubke, C.R., Miller, P., Parker, M.A. and Espejo, E. (2001). Fair play? Violence, gender, and race in video games.

Gonzalez, A., Gomez, E., Orozco, R., Jacobs, S. (2014). Entering the Boys' Club: An Analysis of Female Representation in Game Industry, Culture, and Design. IConference 2014 Proc.

Gray, K.L., Buyukozturk, B. and Hill, Z.G. (2017). Blurring the boundaries: Using Gamergate to examine "real" and symbolic violence against women in contemporary gaming culture. Sociology compass, 11(3).

Hart, C. (2017). Getting into the game: an examination of player personality projection in videogame avatars. Game Studies vol.17(2).

Heritage, F. (2022). Magical women: Representations of female characters in the Witcher video game series. Discourse, Context & Media, 49.

Huh, S. and Williams, D. (2010). Dude looks like a lady: Gender swapping in an online game. Online worlds: Convergence of the real and the virtual, 161-174.

International Game Developers Association (IGDA). (2022). Developer Satisfaction Survey 2021 Diversity in the Game Industry Report.

Jansz, J. and Martis, R.G. (2007). The Lara phenomenon: Powerful female characters in video games. Sex roles, 56, 141-148.

Kennedy, H.W. (2002). Lara Croft: Feminist icon or cyberbimbo?. Game studies, 2(2), 1-12.

Kennedy, H.W. (2018). Game jam as feminist methodology: The affective labors of intervention in the ludic economy. Games and Culture, 13(7), 708-727.

Klastrup, L. and Tosca, S. (2009). "Because it just looks cool!" - Fashion as character performance: The Case of WoW. Journal For Virtual Worlds Research 1(2): 1–17.

Koh, C.J. and Das, R. (2017). Female preferences in video games: a literature review. Gender in Science & Technology, 20.

Kondrat, X. (2015). Gender and video games: How is female gender generally represented in various genres of video games?. Journal of comparative research in anthropology and sociology, 6(01), 171-193.

Lankoski, P. (2002). Character design fundamentals for role-playing games. Beyond Role and Play, 139-148.

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A.R., Evans, C., and Vitak, J. (2008). Teens, Video

Games, and Civics: Teens' Gaming Experiences Are Diverse and Include Significant Social Interaction and Civic Engagement. Pew internet & American life project.

Lenhart, A., Smith, A., Anderson, M., Duggan, M., and Perrin, A. (2015). Teens, Technology & Friendships: Video games, social media and mobile phones play an integral role in how teens meet and interact with friends" Pew Research Center.

Leonard, D. (2003). Live in your world, play in ours: Race, video games, and consuming the other. Studies in media & information literacy education, 3(4), 1-9.

Leonhardt, M. and Overå, S. (2021). Are there differences in video gaming and use of social media among boys and girls?—A mixed methods approach. International journal of environmental research and public health, 18(11).

Liu, H. (2018). Early adolescents' perceptions and attitudes towards gender representations in video games.

Lucas, R. (2019). Out of the Margins: Evolving Narrative Representation of Women in Video Games.

Lynch, T., Tompkins, J.E., Van Driel, I.I. and Fritz, N. (2016). Sexy, strong, and secondary: A content analysis of female characters in video games across 31 years. Journal of Communication, 66(4), 564-584.

MacCallum-Stewart, E. (2014). 'Take That, Bitches!'Refiguring Lara Croft in Feminist Game Narratives. Game Studies, 14(2).

Maclean, E. (2016). Girls, guys, and games: How news media perpetuate stereotypes of male and female gamers. Press Start, 3(1), 17-45.

Massanari, A. (2017). # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. New media & society, 19(3), 329-346.

Masso, I.C. (2011). The grips of fantasy: The construction of female characters in and beyond virtual game worlds. In Creating Second Lives (113-142). Routledge.

McArthur, V. (2020). Damsel in this Dress: an analysis of the character designs of women in post-secondary game design programs. Feminist Media Studies, 21(3), 381-397.

Miller, M.K. and Summers, A. (2007). Gender differences in video game characters' roles, appearances, and attire as portrayed in video game magazines. Sex roles, 57, 733-742.

Mortensen, T.E. (2018). Anger, fear, and games: The long event of# GamerGate. Games and Culture, 13(8), 787-806.

Mou, Y. and Peng, W. (2009). Gender and racial stereotypes in popular video games. Handbook of research on effective electronic gaming in education, 922-937.

Nakamura, L. (2019). Gender and race in the gaming world. Society and the Internet: How networks of information and communication are changing our lives, 127-45.

Nguyen, T., Khan, N., Berardi, A., Corvite, S., Norman, C., Gardner, D. (2020). Exploring the perceptions of race on video game covers. IConference 2020 Proc.

Ochsner, A. (2017). Reasons why: Examining the experience of women in games 140 characters at a time. Games and Culture, 14(5), 523-542.

O'Leary, A. (2012). In virtual play, sex harassment is all too real. The New York Times, 2.

Passmore, C.J., Yates, R., Birk, M.V. and Mandryk, R.L. (2017). Racial diversity in indie games: Patterns, challenges, and opportunities. In Extended abstracts publication of the annual symposium on computer-human interaction in play, 137-151.

Perreault, M.F., Perreault, G.P., Jenkins, J. and Morrison, A. (2018). Depictions of female protagonists in digital games: A narrative analysis of 2013 dice award-winning digital games. Games and Culture, 13(8), 843-860.

Phan, M.H., Jardina, J.R., Hoyle, S. and Chaparro, B.S. (2012). Examining the role of gender in video game usage, preference, and behavior. In: Proceedings of the human factors and ergonomics society annual meeting . Sage CA: Los Angeles, CA: Sage Publications. 56(1), 1496-1500.

Qin, H., Patrick Rau, P.L. and Salvendy, G. (2009). Measuring player immersion in the computer game narrative. Intl. Journal of Human–Computer Interaction, 25(2), 107-133.

Reza, A., Chu, S., Nedd, A. and Gardner, D. (2022). Having skin in the game: How players purchase representation in games. Convergence, 28(6), 1621-1642.

Ruberg, B., Cullen, A.L. and Brewster, K. (2019). Nothing but a "titty streamer": legitimacy, labor, and the debate over women's breasts in video game live streaming. Critical Studies in Media Communication, 36(5), 466-481.

Saldaña, J. (2015). The Coding Manual for Qualitative Researchers. Sage Publications Ltd.

Salter, A., Blodgett, B. (2017). Come get some: damsels in distress and the male default avatar in video games. Toxic Geek Masculinity in Media: Sexism, Trolling, and Identity Policing, 73-99.

Scharrer, E. (2004). Virtual violence: Gender and aggression in video game advertisements. Mass Communication & Society, 7(4), 393-412.

Shaw, A. (2009). Putting the gay in games: Cultural production and GLBT content in video games. Games and Culture, 4(3), 228-253.

Shaw, A. (2014). Gaming at the edge: Sexuality and gender at the margins of gamer culture. U of Minnesota Press.

Shaw, A. (2017). "Diversity without defense: Reframing arguments for diversity in games." In Kinephanos: Journal of Media Studies and Popular Culture, July, 54-76.

Shaw, A. and Friesem, E. (2016). Where is the queerness in games?: Types of lesbian, gay, bisexual, transgender, and queer content in digital games. International Journal of Communication, 10, 13.

Sherman, S.R. (1997). Perils of the princess: Gender and genre in video games. Western folklore, 56(3/4), 243-258.

Shliakhovchuk, E. and Muñoz García, A. (2020). Intercultural Perspective on Impact of Video Games on Players: Insights from a Systematic Review of Recent Literature. Educational Sciences: Theory and Practice, 20(1), 40-58.

Solska, D. (2022). The evolving female narrative in dystopian video games: the case of BioShock Infinite, The Last of Us, and Horizon: Zero Dawn. Currents, 7.

Stermer, S.P. and Burkley, M. (2015). SeX-Box: Exposure to sexist video games predicts benevolent sexism. Psychology of Popular Media Culture, 4(1), 47.

Summers, A., and Miller, M.K. (2014). From damsels in distress to sexy superheroes: How the portrayal of sexism in video game magazines has changed in the last twenty years. Feminist Media Studies, 14(6), 1028-1040.

Super Mario Wiki. (n.d.). List of Princes Peach quotes, https://www.mariowiki.com/List_of_Princess_Peach_quotes.

Tomkinson, S.: "She's Built Like a Tank". (2023). Player Reaction to Abby Anderson in The Last of Us: Part II. Games and Culture, 18(5), 684-701.

Tompkins, J.E., Lynch, T., Van Driel, I.I. and Fritz, N. (2020). Kawaii killers and femme fatales: A textual analysis of female characters signifying benevolent and hostile sexism in video games. Journal of Broadcasting & Electronic Media, 64(2), 236-254.

Tompkins, J.E. and Martins, N. (2021). Masculine pleasures as normalized practices: Character design in the video game industry. Games and Culture, 17(3), 399-420.

Ugolotti, V.C., Brusk, J. and Engström, H. (2020). Crying in the Game-An inductive analysis of a game that portraits benevolent sexism. In: Proceedings of the 14th International Conference on Game Based Learning, ECGBL.

Unknown creator. (n.d.). Video Game Logic - Protagonists, https://knowyourmeme.com/photos/787777-video-game-logic.

Weber, R. (2012). 1ReasonWhy-The night Twitter took on the industry's sexism. GamesIndustry.

Williams, D. (2003). The video game lightning rod. Information Communication & Society, 6(4), 523-550.

Williams, D., Consalvo, M., Caplan, S. and Yee, N. (2009). Looking for gender: Gender roles and behaviors among online gamers. Journal of communication, 59(4), 700-725.

Wingfield, N. (2014). Feminist critics of video games facing threats in 'GamerGate'campaign. The New York Times, 15.

Wirtz, B. and Gamedesigning.org. (n.d.). The Issue of Diversity in gaming, https://www.gamedesigning.org/gaming/diversity/.

Wood, J.R. and Wood, L.E. (2008). Card sorting: current practices and beyond. Journal of Usability Studies, 4(1), 1-6.

Yao, M.Z., Mahood, C. and Linz, D. (2010). Sexual priming, gender stereotyping, and likelihood to sexually harass: Examining the cognitive effects of playing a sexually-explicit video game. Sex roles, 62, 77-88.

Zippia. (n.d.). Video game developer demographics and statistics in the US, https://www.zippia.com/video-game-developer-jobs/demographics/.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

State literacy: lessons from social struggle in the Colombian Amazon

Claudia Grisales-Bohórquez

DOI: https://doi.org/10.47989/ir30iConf46930

Abstract

Introduction. This paper introduces the concept of state literacy, drawing on the struggles of *campesino* or peasant communities in the Colombian Amazon and interdisciplinary literature on state-citizen interactions.

Method. Using feminist and action-research methodologies, the study is based on ethnographic fieldwork in Guaviare, Colombia (April–December 2023). Data collection involved participant observation, interviews with peasant community leaders and members, and analysis of documents and photographs. Thematic analysis explored campesino interactions with state bureaucracies and their information practices.

Analysis. Historically marginalized by state zoning policies that deem their lands unsuitable for peasant practices, campesino communities have organized collective efforts to reinterpret state governance frameworks. They use tools such as state maps and digital applications to foster state literacy, challenging exclusionary policies and advocating for alternative governance models.

Results. State literacy emerges as a collective, networked, and critical practice that enables marginalized communities to engage with state systems. It is supported by informal sociotechnical infrastructures. It configures subversive practices that transform bureaucratic systems.

Conclusions. State literacy represents a vital set of skills and practices for marginalized communities to navigate and resist bureaucratic marginalization. It highlights the potential of marginalized communities to challenge and reimagine governance structures, while proposing ethical, community-driven alternatives.

Introduction

While working with *campesino*, or peasant, community leaders in the northwestern border of the Colombian Amazon, I was struck by their deep familiarity with the department's land-planning terminology. Terms like '*forest preserve zone Type A or B*' were part of everyday conversations among them. People responded to the question '*what figure is your farm located in*?' as naturally as I learned to answer questions about my ethnicity in surveys during my first year as an international student in Illinois. However, while my categorization as '*Hispanic*' has been a minor inconvenience at most, for my interlocutors, living in Type A or Type B areas of the forest preserve zone entails a specific governance regime that limits access to rights and the realization of family and community projects.

As I continued working with peasant community leaders, my attention was continuously drawn to their knowledge of the state—how they developed and used it, and why. Perhaps because my main research interest was on the integration of information technologies to peasant struggles in the Amazon, I noticed that digital information technologies played a key role in supporting this knowledge. Drawing on this modest insight, this short paper introduces the concept of state literacy, defined as the ability to obtain, understand, and engage with information about the conditions of governance of one's community in situated encounters with the state. By looking at the illustrative case of one particular peasant community in the department of Guaviare, I identify three characteristics of state literacy emerging in the struggles of marginalized communities: 1) it is collective, networked, and critical, 2) it is embedded in alternative sociotechnical infrastructures, and 3) it configures subversive information practices. Mobile phones have become key tools for peasant community leaders, and they are part of each of these characteristics.

Background literature

Researchers have centered the agency of marginalized communities and individuals, examining their responses to barriers and risks imposed by structural inequalities (Espinoza Vasquez & Oltmann, 2023; Kitzie et al., 2020). Espinoza Vasquez and Oltmann identify '*defensive, proactive, and subversive information practices*' employed by communities to navigate information marginalization. Defensive information practices are those in which people react to new barriers, seeking to protect themselves, their family, or their community. Proactive information practices are actions undertaken in anticipation of actual barriers or risks, stemming from social interdependency and shared vulnerability and deep understanding of community needs, structure, and knowledge. Subversive information practices are actions that counter or challenge the status quo and systemic inequalities (Espinoza Vasquez & Oltmann, 2023, p. 528).

Within science and technology studies (STS), '*material literacy*' describes how communities navigate bureaucratic infrastructures by sharing tacit knowledge about their material constraints. This knowledge helps create alternative infrastructures, enabling new forms of existence and restoring justice through alliances between committed groups of humans and accessible material objects (Mora-Gámez, 2020; Papadopoulos, 2018). In these situations, Espinoza Vasquez (2023) identifies activists' '*seamful work*' (Vertesi, 2014), where situated knowledge and available resources (actors, organizations, technologies) are used to build sociotechnical infrastructures that counter state-driven inequalities. Research has also shown how communities organize to intervene state infrastructures and tools of governance that '*quietly determine the limits of the possible by both narrowing down certain options and opening the possibility of creating different, and maybe better, worlds*' (Ballestero, 2019, p. 5). This type of activism requires a nuanced understanding of the relationship between knowledge/legibility and power, as explained by scholars like Foucault or Scott (Foucault, 2012; Scott, 2008).

In public administration and political science, scholars have explored everyday encounters between subaltern citizens and the state, focusing on how the state's failure to deliver services

disproportionately affects those lacking cultural capital, bureaucratic literacy, or social networks (Moodie, 2013). Döring (2021) extends the concept of health literacy to 'administrative literacy', defined as '*the capacity to obtain, process, and understand basic information and services from public organizations needed to make appropriate decisions*' (p. 1155). Administrative literacy often emerges in specific contexts over time, as citizens compensate for the state's inability to provide training opportunities or usable information (Abou Moumouni & Krauß, 2023). Moodie (2013) illustrates how bureaucratic literacy gained through state employment may become collective knowledge, supporting community struggles through imaginative applications. Her research invites attention to the importance of bureaucratic literacy emerging from unintended information flows and proximities in state policy.

Research in community informatics has long explored the relationship between information infrastructures and the practices of social movements and local communities. For example, Jack Linchuan Qiu introduced the concept of the *'have-less'* to describe social groups in China who had limited access to ICTs but used them to network their communities under structural constraints (2009). Similarly, David Nemer studied how marginalized residents of Brazil's favelas appropriated technology centers to navigate oppressive conditions (2022). In Colombia, Rocio Rueda Ortiz and her team analysed how an indigenous organization integrated digital technologies into their struggles by centering the life and work of one of the organizations' key figures (Rueda Ortiz, 2011). Researchers have also examined the domain learning that occurs in interactions with digital infrastructures, which is often mediated by social actors that help others achieve certain tasks online (Prieto-Nanez, 2016; Williams, 2012). Finally, they have studied forms of community learning and teaching mediated by computing infrastructures (Grisales Bohorquez et al., 2021; Levy, 1999).

Method

This research draws on ethnographic fieldwork conducted between April and December 2023 in Guaviare, Colombia. Working with eight campesino leaders, I employed feminist and action-research methodologies, conducting participant observation, collecting field notes, documents, and photographs, and focusing on information practices and interactions with state bureaucracies. Semi-structured interviews were conducted with seven community leaders and eight community members to explore their relationships with information technologies and discuss ethnographic observations. The data was analyzed using thematic analysis and iterative coding, emphasizing the work and knowledge campesinos use to navigate and challenge marginalization.

Results

In 2013, the Colombian state divided the lands of the forest preserve zone in Guaviare into Type A areas, reserved for ecological processes, and Type B areas, designated for sustainable forestry management. The zoning effectively rendered peasants *'unfit'* to inhabit these lands, reinforcing stereotypes of peasants as coca-growers allied with armed guerrillas. In response, peasant communities in the municipality of Calamar, Guaviare have organized to demand recognition as valid interlocutors in territorial management and access to rights. For more than a decade, they have engaged in protective, proactive, and subversive practices, including refusing to sign documents and provide information, writing letters, attending meetings and trainings with state officer and NGOs, participating in mobilizations, sought the support of lawyers and activists, and collectively developed strategies to achieve the recognition of community expectations and plans for their own future. In 2023, community leaders presented a comprehensive proposal for the recognition of their collective land rights that was grounded in existing Colombian law, but alternative to the state's proposed strategy. This demonstrated an understanding of conditions under which the state made peasants and their land legible, and thus governed or administered them (Scott, 2008). The community's proposal made peasants legible as environmental agents, demanding recognition of their collective rights inhabit and decide over their territory as a strategy for the protection of the Amazon jungle based on community environmental governance.

State literacy in practice

In ongoing interactions with state officers, documents and spaces, community leaders gain information about how the state operates (or not). Miguel, a community leader, recounted how he gained access to a crucial piece of information–a state map used to delineate environmental zones:

> *those maps were used by professionals from the corporation or the environmental ministry to identify where environmental damage occurred, for example. I became friends with one and convinced him to give me the map, though he warned me not to reveal the source. This tool, though informally obtained, has helped us resolve conflicts with environmental authorities. Before, we were always defensive because we didn't understand the zoning, we didn't even know where we were standing. Now, this tool is crucial. It has given us elements of weight for discussions with the institutions and allowed us to make more grounded proposals in relation to the environmental zoning of the territory* (community leader, 42, personal communication).

Using a geospatial PDF file on the Avenza maps app that he got from his friend, Miguel shared state perspectives on their lands with his community, fostering understanding of the state's environmental priorities and revealing peasants' exclusion as obstacles to conservation. In this story, Miguel's experience illustrates *state literacy*–the ability to obtain, understand and engage with information about the conditions of governance of one's community. State literacy, supported by a heterogeneous socio-technical network, empowered campesino leaders to continue to challenge state narratives and propose grounded, alternative governance models.

Characteristics of state literacy

The following three characteristics of state literacy were identified in the struggles of peasant communities in Calamar, Guaviare.

Collective, networked, and critical

State literacy of peasant communities in Calamar is an ongoing collective achievement. Every new encounter with state bureaucracies provides information about old or new constraints that is discussed with neighbors in everyday conversation, in community WhatsApp groups and in community assemblies, particularly when they result in failure to receive attention. Additionally, individual skills and knowledges are embedded in community practices of consensus and reciprocity. While individuals cultivate specific skills used for personal benefit, quotidian information practices are used to guide individual literacies toward the common good, including making jokes about individualist behaviors and asking for help in navigating state bureaucracies. As state interactions have become increasingly digitized, neighbors have helped each other learn and use digital technologies for these purposes.

State literacy is networked because it relies on distributed knowledges and social mediators working from the local to the international scale. In dealing with the state, it is true that nobody knows everything, everyone knows something, and all knowledge resides in networks, paraphrasing Pierre Levy (1999). Peasant leaders recurrently call on their weak links; national activists, NGOs, and ally state officers, whose sporadic interactions with members of remote local communities are extended through networked technologies.

Finally, it is critical because it emerges from accumulated experiences of information marginalization. Critical state literacy enables communities to engage with and transform bureaucratic systems, viewing them as dynamic and malleable rather than fixed or determined structures.

Embedded in alternative sociotechnical infrastructures

State literacy of this marginalized community is sustained by heterogeneous sociotechnical infrastructures that community leaders patch together to address information asymmetries and

barriers. For example, Miguel combines a state map, situated knowledge, a smartphone, community assemblies and casual conversations to build informal infrastructures that disseminate state literacy and sustain collective practices. Smartphones and internet infrastructures, however limited, have been key elements of peasants' alternative sociotechnical infrastructures since 2016, when state and international cooperation agencies began implementing projects in Calamar to support the peace agreements with armed guerillas. Smartphones, and specifically WhatsApp, began mediating the relationships between local community leaders and institutions, whereas community leaders were the designated as information brokers by their communities. This is why most of them gave a similar account: '*when I became president, leader, the community made me buy a phone.*'

Configures subversive information practices

State literacy is heightened by subversive information practices that destabilize the status quo and challenge inequalities, and it contributes to subversive practices of marginalized communities. Miguel's appropriation of the map exemplifies a subversive information practice that challenges information asymmetries framing peasants as ignorant of environmental zoning. Using his smartphone, Miguel shared state literacy stemming from this subversive practice with his community. Collectively, campesinos used this critical state literacy in subversive information practices that challenge governance structures while proposing alternative models prioritizing conservation and campesino rights as commentary rather than contradictory. For example, in 2023 groups of peasants georeferenced the borders of the collective territory they demanded recognition for, using GPS devices and smartphones, as well as community agreements regarding where the borders should go. The information was patched together by allied experts from a national NGO and a state agency to create a new state map of the territory including a patch of land for peasants *and* conservation, which challenged stereotypes and institutional constraints to peasants' recognition as environmental agents.

Conclusions

This paper has introduced the concept of state literacy, demonstrating how it emerges from the struggles of marginalized communities as a set of skills, knowledge, and attitudes enabling communities to navigate, challenge, and transform their conditions of governance. Future research should explore how peasant leaders carefully resist governance by illegal armed groups while simultaneously dealing with state marginalization, a dangerous endeavor given the ongoing threats to social leaders in Colombia. Since 2016, over 1,200 social leaders and human rights defenders have been murdered in the country (Human Rights Watch, 2023). This situation highlights the importance of supporting community leaders and human rights defenders by approaching '*access to information*' from their situated perspectives, considering the specific risks, burdens, and barriers they face and working with them to address them. Further research should also look at the emergence and use of state literacy in the struggles of marginalized communities in diverse situations, so that a more general framework related to theory of information marginalization and precarity might be proposed.

Acknowledgements

I am grateful for the direct and indirect contributions of all community members in Calamar, Guaviare, especially of the social leaders and neighbours that sustained and enabled my research with their generosity and encouragement. I acknowledge the support of the *Centro de Alternativas al Desarrollo* (CEALDES) in Colombia, whose sustained work in Guaviare since 2016 allowed my research to contribute to generative political alliances between "technical experts" and scholar, peasant communities and the Amazon jungle. I extend my appreciation to the anonymous referees that helped improve my contribution.

About the author

Claudia Grisales Bohórquez is a Ph.D. candidate in Informatics at the University of Illinois, Urbana-Champaign, USA. Her research interests are in sociotechnical and technopolitical practices of marginalized communities and community appropriation of information technologies. She can be contacted at clg3@illinois.edu.

References

Acuña, S. T., Castro, J. W., & Juristo, N. (2012). A HCI technique for improving requirements elicitation. Information and Software Technology, 54(12), 1357-1375. https://doi.org/10.1016/j.infsof.2012.07.011

Andronico, P., Buzzi, M., Castillo, C., & Leporini, B. (2006). Improving search engine interfaces for blind users: a case study. Universal Access in the Information Society, 5(1), 23-40. https://doi.org/10.1007/s10209-006-0022-3

Ballestero, A. (2019). *A future history of water*. Duke University Press.

Bannon, L. (2011). Reimagining HCI: toward a more human-centered perspective. Interactions, 18(4), 50–57. https://doi.org/10.1145/1978822.1978833

Abou Moumouni, I., & Krauß, R. (2023). 'Paperwork is so important!': Processes of literacising in bureaucratic contexts in Benin and Bolivia. *Cultura & Psyché*, 4(1), 85–106. https://doi.org/10.1007/s43638-023-00085-z

Döring, M. (2021). How-to Bureaucracy: A Concept of Citizens' Administrative Literacy. *Administration & Society*, 53(8), 1155–1177. https://doi.org/10.1177/0095399721995460

Espinoza Vasquez, F., & Oltmann, S. M. (2023). Information precarity and the agentic practices of marginalized communities: Puerto Rican activists addressing the crisis before, during, and after Hurricane Maria. *Journal of the Association for Information Science and Technology*, 74(5), 517–530. https://doi.org/10.1002/asi.24742

Foucault, M. (2012). *The order of things*. Taylor & Francis. http://public.eblib.com/choice/publicfullrecord.aspx?p=240649

Grisales Bohorquez, C., Kay, M., Lenstra, N., & Williams, K. (2021). Age and Technology in Action. In W. Aspray (Ed.), *Information Issues for Older Americans*. Rowman & Littlefield Publishers.

Human Rights Watch. (2023). Colombia: Eventos de 2023. In *Informe Mundial 2024*. https://www.hrw.org/es/world-report/2024/country-chapters/colombia

Kitzie, V. L., Wagner, T. L., & Vera, A. N. (2020). 'In the beginning, it was little whispers…now, we're almost a roar': 15th International Conference on Sustainable Digital Communities, iConference 2020. *Sustainable Digital Communities - 15th International Conference, iConference 2020, Proceedings*, 15–31. https://doi.org/10.1007/978-3-030-43687-2_2

Levy, P. (1999). *Collective Intelligence*. Basic Books.

Moodie, M. (2013). Bureaucratic Literacy and the Politics of Complaint: Untold Story from Rajasthan. *Economic and Political Weekly*, 48(45/46), 71–75.

Mora-Gámez, F. (2020). Beyond citizenship: The material politics of alternative infrastructures. *Citizenship Studies*, 24(5), 696–711. https://doi.org/10.1080/13621025.2020.1784648

Nemer, D. (2022). *Technology of the Oppressed: Inequity and the Digital Mundane in Favelas of Brazil*. https://doi.org/10.7551/mitpress/14122.001.0001

Papadopoulos, D. (2018). *Experimental practice: Technoscience, alterontologies, and more-than-social movements*. Duke University Press.

Prieto-Nanez, F. M. (2016). Learning Through Infrastructures: Cybercafes as Spaces for Digital Literacy. In M. A. Peters (Ed.), *Encyclopedia of Educational Philosophy and Theory* (pp. 1–5). Springer. https://doi.org/10.1007/978-981-287-532-7_473-1

Qiu, J. L. (2009). *Working-class network society: Communication technology and the information have-less in urban China*. MIT Press.

Rueda Ortiz, R. (2011). Cultura, política y repertorios tecnológicos: El caso del tejido de comunicación NASA-ACIN. In *Tierra y silicio: Cómo la palabra y la acción política de pueblos indígenas cultivan entornos digitales*. Universidad del Valle, Programa Editorial.

Scott, J. C. (2008). Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. In *Seeing Like a State*. Yale University Press. https://doi.org/10.12987/9780300128789

Vasquez, F. K. E. (2023). Working at the Seams of Colonial Structures: Alternative Sociotechnical Infrastructures Revealed by Hurricane Maria. *Science, Technology, & Human Values*, 48(2), 374–400. https://doi.org/10.1177/01622439211058448

Vertesi, J. (2014). Seamful Spaces: Heterogeneous Infrastructures in Interaction. *Science, Technology, & Human Values*, 39(2), 264–284. https://doi.org/10.1177/0162243913516012

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Analysis of the integration of e-commerce and rural lives in China

Guangchun Zheng, Sharifa Sultana, and Kate Williams

DOI: https://doi.org/10.47989/ir30iConf46927

Abstract

Goal. The aim of the study is to investigate on how e-commerce is integrated in Chinese rural farmers' lives and its impact.

Data. The primary data consists of 30 social media posts from four popular social media e-commerce platforms in China: Weibo, Xiaohongshu, TikTok, and Kuaishou.

Analysis. The study used the performance of rurality with computing as theoretical lens to conduct content analysis. The posts were first coded under three themes: infrastructure, situated use, and geography, and then summarized into keywords.

Findings. The study found out that the current state of rural e-commerce in China validates the uniqueness and richness of rurality. The vibrant vitality of rural e-commerce proves that rural areas are not passive recipients of assistance but can be active agents of development.

Contribution. This study contributes to ICT4D research by providing implications for both technological design of e-commerce platforms and rural revitalization.

Introduction

E-commerce has been developing fast in China over the past years, and rural areas account for a good portion of it. A survey in 2012 showed that residents in smaller and more remote cities spent over 20% of their disposable income online, which exceeds the proportion among residents in China's largest cities (Fan et al., 2018). In recent years, more information and communication technologies (ICTs) like live streaming have been integrated with agriculture and facilitate local development.

Scholars believe rural e-commerce is an opportunity to boost social innovation and social welfare (Huang et al., 2020), develop the local economy (Liu et al., 2021), and realize poverty alleviation (Peng et al., 2021). Instead of objects in need of assistance, rural areas are the agencies which can help themselves. This community-driven development model, which is often generated from the bottom up, demonstrates rurality as a unique resource that can provide insights on e-commerce development models in other regions as well as technology design. Therefore, this study seeks to answer the following research questions: 1) How is e-commerce integrated in Chinese rural farmers' lives? 2)What impact has e-commerce use had on rural and urban areas?

Background and literature review

E-commerce and rural e-commerce in China

The development of e-commerce in China began in the late 1990s. At the beginning of the 21st century, Taobao and JD.com, which played pivotal roles in China's e-commerce development, were established successively (Huang & Zhou, 2015). Not all scholars are optimistic about China's emerging market. In his work published in 2008, Martinsons argued that limited by a comparatively traditional business model characterized by 'personal trust, contextual and informal information,' China will face unique challenges. However, Chinese e-commerce has ushered in a period of sustained and stable development since 2010. By 2013, China had become the world's largest e-commerce market (Kwak et al., 2019).

China's rural e-commerce developed along with the overall e-commerce industry. The first online retailer of agricultural products, Yiguo Fresh, was launched in 2005 (Liu & Shen, 2019). Rural access to the Internet provided a foundation for rural e-commerce. China's Eleventh Five-Year Plan (2006-2010) proposed a rural communications development goal in which '*every village has access to the telephone and Internet*' (Central Government Portal, 2011). Rural e-commerce entered its booming stage during the 2010s. In 2022, China's overarching act on agriculture and rural areas proposed '*implementing the "digital business to promote agriculture" project and promoting e-commerce into rural areas*' (China Economic Net, 2022).

Recent literature has described multiple methods to investigate the status quo and impacts of rural e-commerce. E-commerce's role in the increase in the income of rural residents is supported by data from quantitative research (Fan et al., 2018; Peng et al., 2021). Several case studies included local context and further investigated on the role of rural e-commerce in social innovation (Zhang et al., 2022; Cui et al., 2017), poverty alleviation (Huang et al., 2020) and women empowerment (Tang et al., 2022). These studies generally affirm its positive impacts, but Yu and Cui (Yu & Cui, 2019) believed that women's empowerment was not guaranteed by socioeconomic enablement due to traditional constraints on women and digital capitalist exploitation.

Rural computing and ICT4D

In the broader research field of ICT for Development (ICT4D), rural computing is a subfield that receives less attention. Some scholars emphasized the importance of considering technology design from the rural standpoint, because rural areas have special needs (Hardy et al., 2019). When technology is designed with urban users in mind, rural users often encounter mismatches between

their needs and the assumptions embedded in the design, which are based on 'scripted users'—the intended user profiles envisioned by designers (Hardy & Lindtner, 2017).

ACM Transactions on computer-human interaction published a special issue on rural computing and HCI in 2021. In the introduction of this special issue (Su et al., 2021), rurality is not seen as static, but rather actively shaped, enacted, and experienced through technology, or what is referred to as '*performed*'. The performance of rurality with computing is summarized in three themes: infrastructure, situated use, and geography. This study will analyse the performance of rurality in the case of e-commerce in rural China based on the themes.

Tang and their colleagues' work is a good example in this field (Tang et al., 2022). It considered both social and technical backgrounds and studied the impact of emerging live streaming technology on women empowerment. The interactivity of social media has brought new challenges and opportunities to e-commerce. This study focuses on the landscape and characteristics of rural e-commerce in China amidst the widespread use of social media platforms. It aims to offer suggestions for the technical design of these platforms and inspire other rural empowerment projects.

Method

This study collected and analysed 30 social media posts from Weibo, Xiaohongshu, TikTok, Kuaishou. These were all popular social media platforms in China with e-commerce technologies embedded. The data collection process included the following steps: 1) Search on the above platforms using the keywords, 助农 (helping farmers), 农村电商 (rural e-commerce), and 三农 (agriculture, rural areas, and farmers). 2) Browse posts published in 2023 and manually determine whether the post is relevant. 3) For selected posts, record the following information: publishing platform, time, format, content, and type of user. Identifiable information was anonymized in the process. The final sample size for this study was ten posts from local users, who lived or worked in rural China; ten posts from outside users; five posts from common media, and five posts from government official accounts. The 30 posts in full can be viewed from the link: https://docs.google.com/spreadsheets/d/1TBL8PXaYvGZK8I3LZPDPvyfV2e583KZl7RaWzJTTh6w/edit?usp=sharing.

The data analysis method used in this study was content analysis. Using the performance of rurality with computing as theoretical lens (Su et al., 2021), the content of the posts was first coded under the three themes (infrastructure, situated use, and geography) and then summarized into keywords (e.g. economic conditions, social stigma, etc.). The entire process of data collection and analysis was conducted manually and in Chinese, the language of the original posts. The findings, along with some examples of the corresponding content, were translated into English and are presented in the next section.

Findings

Infrastructure

Technical infrastructure

The technical infrastructure of rural e-commerce is mainly built in three aspects: communication network, e-commerce platform, and logistics and transportation. China has made great progress in developing these three aspects. Many towns and villages have almost achieved 100% 4G/5G signal coverage. Easy access to the internet and the increase in smartphone penetration have given rural residents more opportunities to access e-commerce platforms.

Although transportation costs remain a significant challenge for rural e-commerce, there have been notable achievements in the basic coverage of logistics and transportation services. As of 2022, China '*has built a total of 990 county-level public distribution centers and 278,000 village-level*

express delivery service stations, and 95% of the country's incorporated villages have achieved express delivery service coverage' (post#4).

Economic infrastructure

The agricultural economy is the main economic sector in rural China. Given China's vast territory, different regions are suitable for cultivating different products. The specialties of certain areas have established brand effects like Gannan oranges. However, agricultural production can be significantly influenced by the natural environment, leading to fluctuations in farmers' incomes. '*The total annual income is about ¥10,000 if (tea) is well grown, and ¥7,800 if not.*' (post#24).

The traditional sales method for farmers involves selling their products to middlemen, who then distribute to supermarkets or factories through offline channels. Many '*agricultural bloggers*' on social media are buyers from other regions. Their role is similar to that of traditional middlemen. The difference is that e-commerce relies more on online channels and enables farmers to not only sell products but also resell to other online stores or directly retail online:

> *the building behind me is my mango warehouse. It is a supply chain to provide clients with one-click delivery… The customers I have supplied to include stores on Tmall, Pinduoduo, WeChat merchants (e-commerce platforms in China), some supermarkets, and some export businesses. Since the short videos and live streaming are becoming popular, now I'm also selling my own mangoes through these channels* (post#14).

Social infrastructure

The aging of the population structure is a common problem faced by rural China. As the younger generation migrates to cities, only the elderly remain in the family to manage agricultural production. They often face physical problems brought about by age and long hours of labour, which make it difficult for them to complete the work.

> *My 79-year-old grandpa picks (the fruits) by himself every day, and his hand was fractured due to picking. He cannot use any strength in his hands, so the baskets (of fruits) have to be carried down the mountain by my grandma* (post#19).

Rural e-commerce also faces the problem of social stigma. One stigma is that rural areas are synonymous with backwardness and poverty. Rural families themselves often regard '*leaving the village*' as the life goal. Young people who go out to study and return to the village may suffer from the stigma of '*learning nothing outside*' (post#11). There is also social stigma on the e-commerce industry. China values getting rich through hard work. Running e-commerce on social media seems to require little effort and may be classified as an opportunistic way of making money:

> *(Some people who do e-commerce) lie at home, make videos, pretend to be weird, talk about weird things (to attract others), then put products on the online stores, ship them with just one click, and make money with no effort* (post#8).

Situated use

From attention to purchase: logic behind social media e-commerce

The e-commerce model based on social media platforms is very different from the traditional e-commerce model. Users who purchase products on social media platforms are usually attracted by the content on the platform first, and then become interested in the products. This purchasing logic directly affects the producers' marketing strategies. Gaining attention on these platforms is crucial to success. Some rural residents use social media for entertainment and to share their lives. While they may not initially intend to engage in rural e-commerce, under the 'attention-purchase' business model, gaining a large following can easily lead them in that direction:

> *She saw that other people around her often watched short videos, so she took videos of her daily life. She unexpectedly became popular. Following the suggestions from online comments, she tried e-commerce to sell goods* (post#2).

Easier or harder: literacy requirements for social media e-commerce

Social media platforms have relatively lower literacy requirements for the usage and have provided more people with opportunities to participate in e-commerce. Compared with text and pictures, short videos and live streaming lower the threshold for information exchange and sharing. However, there is a gap between recording life and achieving commercial success. Being able to use software is just the most basic skill. How can one craft compelling content to entice clicks and encourage thorough reading? How can one take attractive product photos? These are skills that ordinary farmers typically lack. This, to some extent, explains why many agricultural influencers are university graduates who return to their hometowns to start businesses, as they help bridge the skill gap:

> *I asked my dad to take some photos as materials for Xiaohongshu posts. He took photos casually. He just took an orange and put it in a bag and take a photo. The photos are not beautiful, but they are indeed the most realistic look of our oranges* (post#12).

Geography

Rural empowerment: increase in income, employment, and cohesion

Online platforms help reduce the intermediate resale links of agricultural products, allowing more profits to reach the hands of farmers. Planting regions that have achieved a certain scale can generate more employment opportunities. A young entrepreneur said that he called people from two villages to help with packaging: '*more than 260 elderly people over the age of 65 have been re-employed in the village, helping to increase their income by more than 100,000 yuan*' (post#5).

The people in the village have a common cause to strive for, work together, and have reaped the rewards; this process has strengthened the cohesion of the community. It can also be a good opportunity for villagers to change their traditional views on the rurality, and they can gain satisfaction and a sense of accomplishment from their contributions to their families and hometowns.

> *Today plums have started to be packaged and shipped again. The whole village got together, and it was almost more lively than the Chinese New Year. We even set off firecrackers when the first load of goods was shipped out* (post#15).

Flow of resources: enhanced communication between urban and rural areas

Rural e-commerce is not a one-way poverty alleviation project. Instead, it is a two-way flow of resources between urban and rural areas. Social media opens a window to showcase rural life, allowing people to see various types of lifestyles. Many urban residents yearn for the idyllic and leisurely pace of life in rural areas, and this window serves as a solace for them. While e-commerce platforms increase income for farmers, they also bring local specialty products to the outside, and even to international markets. This exchange of resources creates a win-win situation. *We don't make profit selling fruits (from the farmers). We just want to ensure that everyone can buy a wide variety of products in our store* (post#18).

Discussion

Consistent with findings in existing literature (Fan et al., 2018; Liu et al., 2021; Peng et al., 2021), this study confirms the positive impact of e-commerce on increasing rural residents' income based on the analysis of rural e-commerce-related posts on four social media platforms. However, whether it can achieve overall empowerment for rural areas, especially in transforming the current social stigmatization, remains a topic for discussion. The study provides some positive evidence, such as

the increased cohesion in rural communities, a shift in the mindset of the younger generation towards rural life and encouraging outcomes in product commercialization and scale expansion by some entrepreneurs. However, attention must be paid to some negative aspects of this issue. The reliance on audience reach for e-commerce on social media platforms has led to many intermediaries creating intentionally sensational and unrealistic videos to gain more attention and sympathy. Such videos not only damage customers' trust in the rural e-commerce industry but also perpetuate stereotypical impressions of rural backwardness.

The findings of this study also shed light on the digital divide existing in rural areas. The existence of the digital divide determines whether rural residents can equally enjoy the benefits brought by e-commerce technology. Drawing from van Dijk's (2005) deconstruction of the digital divide, access to ICT can be viewed as physical access, motivational access, skill access, and usage access. Physical access is no longer the main source of the gap with the vigorous promotion of information infrastructure by the government. From the design perspective of social media platforms, founders utilize short videos and live streaming as mediums to facilitate information exchange, lowering skill requirements and creating a more inclusive technological environment. Furthermore, the combination of entertainment and commerce in social media e-commerce, at a motivational level, encourages rural users to engage with new technologies. Finally, in terms of actual effectiveness, we can observe that groups originally in socially disadvantaged groups have improved their lives through e-commerce technology. For example, two rural residents with disabilities who couldn't work regularly gained the ability to earn money through live streaming sales, increasing their senses of success. All the evidence above shows that the integration of e-commerce with rural lives has the potential to positively impact bridging the digital divide.

Conclusion

By analyzing the experiences and insights shared on social media by rural e-commerce practitioners, both those living in rural areas and elsewhere, combined with external observations from common media and the government, this study presents the fundamental landscape of the integration of e-commerce with rural lives in China. The relatively comprehensive technical infrastructure, an agriculture-centered economic infrastructure, and specific challenges in social infrastructure, collectively shape the distinctive appearance of rural e-commerce in China. Furthermore, the results demonstrate the similarities and differences between the interpretations of technology by platform designers and users, providing inspiration for the design of social media e-commerce platforms. Finally, the vibrant vitality of rural e-commerce proves that rural areas are not passive recipients of assistance but can be active agents of development. Rural areas are not isolated; they maintain a flow of resources with urban areas.

Rural e-commerce is a technology still in development, and future research could explore ways to better utilize it. This study identifies two promising entry points: technology adoption driven by policies aimed at bridging the digital divide and intergenerational support facilitated by young people returning to their hometowns. While the study is rooted in the specific context of rural China, the significance of rural e-commerce for rural revitalization offers valuable insights that could be applied to the development of other regions worldwide.

Acknowledgements

This research is initially an assignment for the course CS598 Computing and Global Development at the University of Illinois, Urbana-Champaign. I would like to thank the instructor Dr. Sharifa Sultana and all my classmates for their comments and suggestions on the structure and data collection of this paper.

About the authors

Guangchun Zheng is a doctoral student in Information Science, University of Illinois at Urbana-Champaign. Her research area is community informatics. Her contact email is gzheng6@illinois.edu

Sharifa Sultana is Assistant Professor in Computer Science, University of Illinois at Urbana-Champaign. She achieved Ph.D. in Information Science at Cornell University. Her research area is human-computer interaction, and human-centered AI. Her contact email is sharifas@illinois.edu

Kate Williams is Associate Professor Emerita in School of Information Science, University of Illinois at Urbana-Champaign. She achieved Ph.D. in Information Science at University of Michigan. Her research focuses on community informatics and public libraries past, present, and future. Her contact email is katewill@illinois.edu

References

Central Government Portal. (2011, January 7). China fully realized 'every village has telephone access, and every township has access to the Internet.' https://www.gov.cn/gzdt/2011-01/07/content_1779987.htm

China Economic Net. (2022, July 8). Rural e-commerce enters a new stage of rural development. https://baijiahao.baidu.com/s?id=1737730274787566015&wfr=spider&for=pc

Cui, M., Pan, S. L., Newell, S., & Cui, L. (2017). Strategy, Resource Orchestration and E-commerce Enabled Social Innovation in Rural China. The Journal of Strategic Information Systems, 26(1), 3–21. https://doi.org/10.1016/j.jsis.2016.10.001

Fan, J., Tang, L., Zhu, W., & Zou, B. (2018). The Alibaba effect: Spatial consumption inequality and the welfare gains from e-commerce. Journal of International Economics, 114, 203–220. https://doi.org/10.1016/j.jinteco.2018.07.002

Hardy, J., & Lindtner, S. (2017). Constructing a Desiring User: Discourse, Rurality, and Design in Location-Based Social Networks. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 13–25. https://doi.org/10.1145/2998181.2998347

Hardy, J., Phelan, C., Vigil-Hayes, M., Su, N. M., Wyche, S., & Sengers, P. (2019). Designing from the rural. Interactions, 26(4), 37–41. https://doi.org/10.1145/3328487

Huang, C.-C., Jin, H., Zhang, J., Zheng, Q., Chen, Y., Cheung, S., & Liu, C. (2020). The effects of an innovative e-commerce poverty alleviation platform on Chinese rural laborer skills development and family well-being. Children and Youth Services Review, 116, 105189. https://doi.org/10.1016/j.childyouth.2020.105189

Huang, X. & Zhou, Y. (2015). History, actuality and development of China's e-commerce from the perspective of Alibaba listing. Commercial Research, (12),11-17.

Kwak, J., Zhang, Y., & Yu, J. (2019). Legitimacy building and e-commerce platform development in China: The experience of Alibaba. Technological Forecasting and Social Change, 139, 115–124. https://doi.org/10.1016/j.techfore.2018.06.038

Liu, J. & Shen, W. (2019). Research on the evolution process and path of rural e-commerce. Journal of Commercial Economics, 19, 123-126.

Liu, M., Min, S., Ma, W., & Liu, T. (2021). The adoption and impact of E-commerce in rural China: Application of an endogenous switching regression model. Journal of Rural Studies, 83, 106–116. https://doi.org/10.1016/j.jrurstud.2021.02.021

Martinsons, M. G. (2008). Relationship-based e-commerce: Theory and evidence from China. Information Systems Journal, 18(4), 331–356. https://doi.org/10.1111/j.1365-2575.2008.00302.x

Peng, C., Ma, B., & Zhang, C. (2021). Poverty alleviation through e-commerce: Village involvement and demonstration policies in rural China. Journal of Integrative Agriculture, 20(4), 998–1011. https://doi.org/10.1016/S2095-3119(20)63422-0

Su, N. M., Hardy, J., Vigil-Hayes, M., Veinot, T., & Comber, R. (2021). Introduction: Performing Rurality with Computing. ACM Transactions on Computer-Human Interaction, 28(3), 1–13. https://doi.org/10.1145/3461832

Tang, N., Tao, L., Wen, B., & Lu, Z. (2022). Dare to Dream, Dare to Livestream: How E-Commerce Livestreaming Empowers Chinese Rural Women. CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3491102.3517634

Van Dijk J A G M. The Deepening Divide: Inequality in the Information Society[M]. Thousand Oaks, CA: Sage Publications, 2005:20-21.

Yu, H., & Cui, L. (2019). China's E-Commerce: Empowering Rural Women? The China Quarterly, 238, 418–437. https://doi.org/10.1017/S0305741018001819

Zhang, Y., Long, H., Ma, L., Tu, S., Li, Y., & Ge, D. (2022). Analysis of rural economic restructuring driven by e-commerce based on the space of flows: The case of Xiaying village in central China. Journal of Rural Studies, 93, 196–209. https://doi.org/10.1016/j.jrurstud.2018.12.001

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Narratives for healing: an exploration of interactions within an online community on navigating bullying trauma

Subhasree Sengupta and Yin Yang

DOI: https://doi.org/10.47989/ir30iConf46924

Abstract

Introduction. Online channels can be powerful vessels offering community bonding and support to cope with bullying-induced trauma. Drawing on the theoretical framework of narrative identity, through this initial exploration, we explore the emergent conversational themes of the subreddit (r/bullying) to understand how trauma is disclosed, its impact on the communal identity and the nature of support provided through such online collectives.

Method. Data was collected using the Reddit API (praw) from the subreddit r/bullying. A random sample of 50 comments was analysed using thematic coding for this initial exploration.

Results. A final set of five themes was developed as an outcome of the analysis. These included a wide array of topics ranging from disclosing deep-rooted resentment and trauma to ways of regaining one's self-esteem to cope with psychological stress and struggle. Some themes also covered strategies and resources to develop endurance and resilience.

Conclusions. These initial insights set the stage for unpacking how narrations of bullying and similar traumatic experiences shape and are shaped by information behavior in mediated contexts (such as online channels), which can hold essential perspectives for conceptualizing the future maintenance and sustenance vision of such virtual spaces.

Introduction.

Bullying and other similar incidents of interpersonal harm can severely injure one's self-worth and impair one's ability to function effectively in society (Vanderbilt & Augustyn, 2010). Bullying is an act of domination and often occurs due to deep-rooted resentment, feelings of jealousy, social stigma or a form of perceived inferiority that subjects the one bullied to torment and a depleted sense of self (Einarsen, 1999). Bullying can cause severe trauma and can culminate in several adverse outcomes, which can prove to be damaging to both individuals and society (Brank et al., 2012). Supporting and understanding the bullied individual and their trauma can be pivotal for personal well-being and readjustment (Craig et al., 2007). Hence, finding and creating channels encouraging those experiencing such duress to disclose their experiences becomes crucial to discerning and curbing Bullying (Zych et al., 2015).

In this regard, several studies have indicated how online forums (such as Reddit) offer varied dimensions of support, solidarity, and empathy and help those experiencing traumatic experiences or struggling with self-expression (Kamarudin et al., 2021; Boettcher, 2021; Sengupta & Semaan, 2015). However, community dynamics (including tensions and manifestations of incivility) and norms of information exchange can dictate the type and extent of disclosure (Bilewicz & Soral, 2020; Martey & Stromer-Galley, 2007). While existing scholarship has widely investigated causes and possible mitigation of various types of bullying behaviors, there remains a gap in understanding how online channels (such as Reddit) provide a forum for discussing and processing the impact of Bullying at the individual and societal level.

To address this gap, understanding the type of conversations that manifest through the discourse patterns of these channels in the context of Bullying can play a crucial role in outlining how such online forums can offer empowerment and healing. The framework of narrative identity is a robust framework that outlines the vital linguistic markers that characterize how people construe life stories and, through that, shape their perceptions of self (McAdams & McLean, 2013). Despite the widespread use of this framework to understand conversational dynamics in several social contexts, an understanding of how this framework applies or can be extended to explore information exchange patterns in technology-mediated channels (online forums) remains to be explored, a gap this exploration addresses. Drawing on this framework, we investigated the discourse patterns of r/bullying, which is a commonly used subreddit that focuses on Bullying and associated support. Since the goal is to explore emergent conversational trends, we selected a sample of 50 conversational units to conduct this initial exploratory analysis. The key questions this present study focuses on are:

1. *How does narrative identity manifest through the conversational exchanges of the chosen subreddit (r/bullying)?*
2. *How do community practices and technical features of the online channel impact the perceived narrative identity unfolding in the discourse of r/bullying?*

The key contributions of this initial study are:

1. Providing a novel application and extension of the theoretical lens of the narrative identity framework to investigate how the narration of traumatic life events unfolds within a technology-mediated context of an online forum,
2. Providing a glimpse into how narrative identity construction in cyberspace is impacted by community norms and the technical features associated with such platforms
3. Providing a novel typology of how Bullying and trauma-linked disclosure patterns emerge and color the conversations in online forums can inform a novel extension to existing scholarship on information behavior in cyberspace.

Related work.

This related work section provides a deeper understanding of narrative identity and its link with traumatic information-sharing patterns. Narrative identity is defined as one's internalized and developing life story, which synthesizes characters, plots, and events in the reconstructed past and imagined future to provide life with purpose, meaning, and unity (McAdams & McLean, 2013; McLean et al., 2018). By constructing life stories, individuals communicate to themselves and others who they are, their experiences, and where they envision their lives heading in the future (McLean et al., 2007). The framework of narrative identity covers critical dimensions of life, such as possessing the agency to change one's life course, as well as an interpersonal layer focusing on the importance of communal care, support and fostering a sense of togetherness (McAdams & McLean, 2013; Adler et al., 2016).

In developing narrative identity, social support (i.e., the material and psychological resources individuals perceive to be available or provided to them when needed) is crucial. Aligned investigations indicate that familial bonds and friendships impact the construction of narrative identity (McLean & Jennings, 2012). Further, de Moor et al. (2021) showed that increasing social support might be associated with the presence of redemption sequences (i.e., reframing a negative experience to convey positive meanings) and self-event connections (i.e., connections between life events and self-concept), two aspects of narratives that are essential for understanding identity development (McLean et al., 2019). Taken together, social support exchanged through conversations with social networks has been found to relate to narrative identity in one way or another, which further motivates this exploration.

Despite its applicability, only a few explorations have investigated how the narrative identity framework manifests through the online landscape's emergent conversational dynamics. These explorations, however, mainly focus on understanding the constituents of online identity as constructed through the narration of life events (Baams et al., 2011; Craig & McInroy, 2014). Synergistic scholarship exploring the use of online forums by LGBTQ+ youth has highlighted an emergent misalignment of real vs. virtual identities, offering insights into how narratives in online spaces act as a way for people to explore and expand their existence, which can have both positive as well as negative ramifications (Emmanuel et al., 2015; Bates et al., 2020). Thus, the mediating impact of technical features and conversational norms that promote or deter the constructed narrative pattern becomes important (Halsema, 2021). While many studies situate identity and its manifestation in the digital context, these are mainly based on self-reported data. Thus, they do not capture how identity negotiations happen through interactions and are constrained or enhanced by community dynamics. Thus, our investigation further builds on these studies by providing a deeper insight into how narrative identity is tied to conversational habits and platform tools, especially in trauma-informed information sharing and support.

Data collection and analysis

Given the premise of this study, we selected a subreddit focused on bullying (r/bullying), which had a similar focus and aimed at the project's vision. Data was collected using the Reddit wrapper (PRAW), which resulted in a set of (990 posts and 6564 comments). Posts here refer to the top-level conversation in a thread that initiates further discussion. The data was pruned to remove any missing data points. While no date and time range for the collection process was specified to allow for a full historical search of retriable data, we report the accessible data based on the API norms, policies, and query limits (Proferes et al., 2021). Table 1 presents further information regarding this dataset.

Number of unique authors	1510
Earliest creation date (Month, Year)	02/2024
Latest creation date (Month, Year)	09/2024
Average score (Posts)	6.74
Average score (comments)	2.07

Table 1: Descriptive characteristics of the entire dataset

To conduct this initial analysis, a random sample of 50 comments was analysed using a thematic coding framework (Williams & Moser, 2019). The authors individually coded the develop the first stage of themes. While coding, each author also noted how the emergent theme aligned or did not align with the narrative identity framework. Initial thematic categories with similar meanings were combined to create or refined based on subsequent discussions and deliberations. For example, the categories of '*offering validation towards redemption*' and '*Redemption in the form of solidarity*' were combined as authors found both categories captured similar linguistic signatures. When mapping constructs from the framework of narrative identity with the emergent themes, the construct most closely aligned with the emergent theme was selected. For example, for a theme on '*despair and hopelessness*,' while both meaning-making and contamination were applicable, contamination was selected given that it aligned best with the overtly hostile tone manifesting in the conversations associated with the theme. The final coding schema had a set of five main themes that best captured the observed conversational patterns of the online forum.

Results.

Next, we describe the five resultant themes from the content analysis.

Exploratory narrative processing through sharing personal anecdotes and experiential wisdom

These conversations allude to personal travails and struggles and often attempt to offer empathy to help others cope with pain and suffering encountered when dealing with trauma while also providing a path for endurance. These conversations thus capture a deep sense of self-processing and self-growth, engaging in catharsis (healing their injuries through bonding with others in their shoes), which aligns with the construct of exploratory narrative processing.

For instance, when commenting on one Reddit user's story of their child being bullied, another user shared a similar experience, laments as they recall painful memories but also offers encouragement by emphasizing the importance of self-preservation as a circumstantially essential survival strategy:

> *…I lost my son to the hands of a school yard bully, and I miss him every damn day!! I believe only fight or punch back if necessary and try and make good decisions. You seem like great parents, so keep up the good work, and I hope things work out for you!!*

Redemption through solidarity and advice on survival tactics

These conversations suggest pathways that bullying survivors need to adopt to cope with a painful memory or harm caused by bullying or traumatic experiences. These include strategies for emotional regulation, self-development strategies, and communication approaches (especially if the matters need to be escalated to authorities). These narrations often capture a stand on how to transform the negative experience induced by the trauma into a positive outcome (by seeking redressal or by resisting), the construct of redemption best aligned with these conversations.

For example, the following quote analyses the psychological processes behind bullying and the relevant strategies to deal with bullying.

> *Schools compete with other schools and the amount of damage popular people can do far outranks what loners can do and this is only ever balanced out when media gets involved from things like mass shootings therefore power dynamics will always favour the popular. Curiously the best Anti-bullying approach is actually to sacrifice your own reputation to drag the bully down with you knowing you have a much smaller climb back up as a result of this…*

The author of this comment explains how bullying is a systemic issue of neglect that creates an uneven power dynamic. The discussion on handling bullying provides strategies that can alter the power dynamic by changing response strategies, thereby causing the bully to change their approach.

Contamination through expressions of hopelessness and resignation

These conversations indicate a lack of self-esteem and a feeling of extreme hopelessness and grief. Most of these, similar to the quote below, express a sense of deep resentment and sorrow due to the emotional injury experienced as part of a traumatic event. Hidden in this type of content is a cry for help and support, which may lead other community members to provide support. The excessive negative outlook and self-loathing that emerge through these conversational themes indicate the alignment with contamination. An example comment in this category stated:

> *I am feeling like a failure. I am heavily depressed. I wanted to join another office, interviewed there but haven't heard back yet. I'm losing my mind. I feel so done.*

Meaning-making through acceptance and developing a strong sense of self.

Posts and comments on r/bullying sometimes depict how a person develops a resistive stance when dealing and negotiating with emotionally scarring events such as bullying. Embedded in these conversations are feelings of acceptance, processing of what all a person has tolerated, and, through this, emerges an emotional shield of strength that demonstrates their ability to withstand and move forward with determination. The narrative approach of processing the impact and grasping the key takeaways from the experience aligns with the meaning-making construct of the narrative identity framework.

An example comment (as mentioned below) highlighted the need to break from dysfunctional familial ties as a necessary process to cope with emotional strains and to regain their self-worth:

> *… When we're children, we have no choice and are stuck with family and classmates who bully us. But when we become adults, we have more choice in our lives and can cut ties with bullies much more easily…you have control in most of these situations as an adult and don't have to tolerate such treatment.*

Balancing communion and agency within the online community

Another type of conversation on r/bullying orients to how community members align themselves with other members through either expressing gratitude or handling tensions that may arise as one tries to embed oneself within the community, trying to negotiate the norms and struggles of finding similar others who resonate and understand their reasons for participation. Thus, these narratives capture identity construction as a balancing act between a developing agency or a voice within the online space while negotiating the norms and existing power structures within the online forum. These conversations are best captured through two sub-themes: (a) **acknowledgment of community** and (b) **navigating community tensions**.

The first sub-theme, **acknowledgment**, is often expressed straightforwardly by expressing honest thankfulness for the online community's support. Here is an instance: '*thank you so much for the support. If anything happens, I'll update.*'

Given the volatility and emotional investment associated with handling traumatic interactions, tensions may arise within the online community, too. The quote below shows an example of a community member discussing how certain behaviors in the community seem offensive or cause discord.

> *First time I said, why are you making fun of that guy who just lost his friend due to some misunderstanding? The second one was when I asked why your Avatar doesn't match your real face. Moreover, someone instantly replied who are you to ask this having no Avatar?*

Discussion and conclusion.

The different emergent themes highlighted how the framework of narrative identity intertwines with traumatic life stories about bullying, weaving together several interconnected nuances of self-identification, development of courage and communal stances to provide solidarity and empowerment. Addressing RQ1, we see how the online community creates its unique voice, which manifests in the distinctive patterns of narrative gleaned. As the themes suggest, we see a certain degree of openness, expressions of anguish, torment, and agony in conjunction with the expectation of support and a manifestation of collective empathy. Akin to the conceptual lens of resilience through resistance, most themes offer strategies for resisting and carefully navigating the traumatic ordeal as coping mechanisms (Rodriguez, 2016). Thereby adding additional dimensions to how narrative identity unfolds in trauma-sharing circumstances. These insights may inform how such channels function, are sustained and are essential for augmenting support to navigate bullying and similar traumatic circumstances.

Addressing RQ2, we see exciting glimpses of how technical features (in the form of auto-moderator contributions) play a role in setting a norm of conversational exchanges in the patterns. Auto-moderator, an automated agent, occasionally posts a standard message that outlines key policies and extended norms of engagement. The engagement of auto moderator can play a critical role in constraining conversations, which may impede open conversations, thereby impacting willingness to participate and associated engagement trends. Further, the conversations that show expressions of gratitude and the discussions on tensions within the online forum also highlight how expressing oneself can become constrained by community standards. However, such regulation and moderation can also help preserve the community from misinformation or from emotionally stirring narratives that disrupt the sanctity of such online forums (Stewart, 2021). When taken together, these observations can impact policy and governance mechanisms of online spaces, especially those dealing with trauma-influenced information sharing.

Future work.

Future work will expand on this initial exploration in multiple ways. Firstly, a deeper and more rigorous conversational analysis will be conducted to quantify narrative trends and present a robust framework on how trauma impacts information-sharing and seeking behaviors, as observed through the microcosm of online channels. Emotions play a key role in narrative framing and thus future inspections will unpack the manifestation of emotion and its linkage with narrative identity. Further, augmenting the textual trends with network analysis to explore the roles of different actors and patterns of reciprocity can provide insights into how online channels function as conduits for mobilizing support and empathy in helping the trauma recovery process. Finally, interviews and self-reported data can add validity and motives for participation, informing design and communal policies on making such spaces open, fair, transparent, and safe, which can be crucial for identity negotiation when navigating trauma.

Acknowledgements
The authors would like to thank the reviewers for their comments and suggestions for improving the arguments presented and for future work.

About the authors
Subhasree Sengupta is an Assistant Professor at Florida State University, holding a joint appointment with the Learning Systems Institute and the School of Information. She completed her PhD from the School of Information Studies at Syracuse University. Her interdisciplinary research focuses on the intersection of human-AI collaboration, responsible use and adoption of AI, AI literacy, STEM workforce development, and the impact of technological innovation on learning and teacher training. She can be contacted at ss24da@fsu.edu

Yin Yang is an Assistant Professor in the School of Communication at Florida State University. She completed her Ph.D. from Pennsylvania State University. Her research examines how media messages affect recipients' beliefs, attitudes, and behaviors in health and environmental contexts, with a focus on narrative persuasion and emotions. She is particularly interested in the messaging efforts to promote awareness of and engagement with overlooked and/or stigmatized issues. She can be contacted at yin.yang@fsu.edu

References
Adler, J. M., Lodi-Smith, J., Philippe, F. L., & Houle, I. (2016). The incremental validity of narrative identity in predicting well-being: A review of the field and recommendations for the future. Personality and Social Psychology Review, 20(2), 142-175.

Baams, L., Jonas, K., Utz, S., Bos, H., & van der Vuurst, L. (2011). Internet use and online social support among same sex attracted individuals of different ages. Computers in Human Behavior, 27, 1820-1827.

Bates, A., Hobman, T., & Bell, B. T. (2020). "Let me do what I please with it... Don't decide my identity for me": LGBTQ+ youth experiences of social media in narrative identity development. Journal of Adolescent Research, 35(1), 51-83.

Boettcher, N. (2021). Studies of depression and anxiety using reddit as a data source: scoping review. JMIR mental health, 8(11), e29487.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. Political Psychology, 41, 3-33.

Brank, E. M., Hoetger, L. A., & Hazen, K. P. (2012). Bullying. Annual Review of Law and Social Science, 8(1), 213-230.

Craig, S. L., & McInroy, L. (2014). You can form a part of yourself online: The influ- ence of new media on identity development and coming out for LGBTQ youth. Journal of Gay & Lesbian Mental Health, 18, 95-109.

Craig, W., Pepler, D., & Blais, J. (2007). Responding to Bullying: What works?. School psychology International, 28(4), 465-477.

de Moor, E. L., Van der Graaff, J., van Doeselaar, L., Klimstra, T. A., & Branje, S. (2021). With a little help from my friends? Perceived friendship quality and narrative identity in adolescence. Journal of Research on Adolescence, 31(2), 384-401.

Einarsen, S. (1999). The nature and causes of Bullying at work. International journal of manpower, 20(1/2), 16-27.

Kamarudin, N. S., Beigi, G., & Liu, H. (2021, August). A study on mental health discussion through Reddit. In 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM) (pp. 637-643). IEEE.

Martey, R. M., & Stromer-Galley, J. (2007). The digital dollhouse: Context and social norms in The Sims Online. Games and Culture, 2(4), 314-334.

McAdams, D. P., & McLean, K. C. (2013). Narrative identity. Current Directions in Psychological Science, 22(3), 233-238.

McLean, K. C., Pasupathi, M., & Pals, J. L. (2007). Selves creating stories creating selves: A process model of self-development. Personality and Social Psychology Review, 11, 262–278.

McLean, K. C., & Jennings, L. E. (2012). Teens telling tales: How maternal and peer audiences support narrative identity development. Journal of Adolescence, 35(6), 1455-1469.

McLean, K. C., Lilgendahl, J. P., Fordham, C., Alpert, E., Marsden, E., Szymanowski, K., & McAdams, D. P. (2018). Identity development in cultural context: The role of deviating from master narratives. Journal of Personality, 86, 631-651.

McLean, K. C., Syed, M., Pasupathi, M., Adler, J. M., Dunlop, W. L., Drustrup, D., Fivush, R., Graci, M. E., Lilgendahl, J. P., Lodi-Smith, J., McAdams, D. P., & McCoy, T. P. (2019). The empirical structure of narrative identity: The initial Big Three. Journal of Personality and Social Psychology, 119, 920–944.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. Social Media+ Society, 7(2), 20563051211019004.

Rodriguez, M. C. G. (2016). "The stories we tell each other": Using technology for resistance and resilience through online narrative communities. In Emotions, technology, and health (pp. 125-147). Academic Press.

Sengupta, S., & Semaan, B. (2021, October). 'a guiding light in a virtual haven': A preliminary analysis of conversations around navigating and repairing self-worth in an online professional community for women. In Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (pp. 158-162).

Stewart, E. (2021). Detecting fake news: Two problems for content moderation. Philosophy & technology, 34(4), 923-940.

Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research. International management review, 15(1), 45-55.

Zych, I., Ortega-Ruiz, R., & Del Rey, R. (2015). Systematic review of theoretical studies on Bullying and cyberbullying: Facts, knowledge, prevention, and intervention. Aggression and violent behavior, 23, 1-21.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Are data papers cited as research data? Preliminary analysis on interdisciplinary data paper citations

Kai Li, Pao-Pei Huang, and Wei Jeng

DOI: https://doi.org/10.47989/ir30iConf46918

Abstract

Introduction. Research data sharing and reuse have become increasingly important in modern science, and data papers represent a new academic publication genre aimed at enhancing the visibility, sharing, and reuse of research data. However, whether citations to data papers reflect actual data reuse remains largely unexplored. This paper presents preliminary findings from a project designed to address this gap.

Method. we conducted a content analysis to manually annotate 437 citation sentences from 309 research articles referencing 50 data papers published in *Data in Brief*, a chief academic journal that only publishes data papers. The data papers were sampled from five knowledge domains based on a paper-level classification system.

Results. Our results show that most citations to all selected data papers (89%) are unrelated to the research data being described in the paper, instead focusing on the research findings or methodologies. This suggests that data papers are being cited similarly to traditional research articles, despite their unique purpose and content.

Conclusion. These findings raise questions about the effectiveness of data papers as representations of research data within the scholarly communication system, as well as their utility in quantitative studies on data reuse.

Introduction

Data papers represent a unique academic genre that primarily serves to document and contextualize research datasets within the broader research landscape (Chavan & Penev, 2011). As Chavan and Penev highlighted, data papers embody the core properties of data publication: availability, documentation, citability, and verification (Chavan & Penev, 2011). They provide comprehensive dataset descriptions, link to the data's location, undergo peer review, and crucially, offer a mechanism for citation credit. Unlike traditional research articles that focus on presenting findings and interpretations, data papers shift the spotlight onto the data itself, offering comprehensive metadata and methodological details. This approach aligns with the growing emphasis on data sharing and open science principles in academia (Kratz & Strasser, 2014). Given its strong relevance to the data-driven research paradigm, this new academic genre has gained traction in recent years, with the number of published data papers surpassing ten thousand and continuing to grow. Some universities and researchers now consider data paper submission a standard practice in data publication (Schöpfel et al., 2020; Thewall, 2020).

The value of data papers extends beyond promoting transparency and reproducibility; they provide an approach for acknowledging the crucial contribution of data to the research process (Gorgolewski et al., 2013). In addition, by offering a citable entity for datasets, data papers also create a pathway for crediting data uses, thereby incentivizing data sharing, research reproducibility, and enhancing the perceived value of research data within the scholarly ecosystem. This addresses a significant challenge identified by Tenopir and colleagues (Tenopir et al., 2011), who noted that the lack of robust reward mechanisms often impedes data sharing activities.

However, despite the increasing prevalence of data journals (such as *Data in Brief* and *Scientific Data*) and data papers, studies indicate that data citation and reuse remain lower than anticipated and that citations to data papers may not equalize data reusing (Jiao & Darch, 2020; Stuart, 2017; Thelwall, 2020). They remain significant challenges to the existing infrastructure to support data sharing and reusing. These discrepancies suggest the presence of factors influencing the impact and integration of data papers in the research system, which are not well understood in the literature. Understanding how data papers are used and cited, particularly the factors behind the citation behavior, is crucial for assessing their effectiveness in promoting data sharing and reuse.

Therefore, our study aims to explore the citation practices surrounding data papers, explicitly focusing on their citation contexts: whether a data paper is cited as research data. This investigation is guided by the overall research interest: What is the role and impact of data papers in scholarly communication, especially from the perspective of the intention of their citations? More specifically, we aim to address:

> RQ1: How are data papers cited in research articles, and for what purposes?
>
> RQ2: How does the context of data paper citations differ between disciplines and citation year gaps?

Through this preliminary analysis, we seek to contribute to the ongoing discussion about the roles of data papers in scholarly communication and their relationship to research data. The findings of this study will have implications for data citation practices, data paper guidelines, and the broader understanding of how data is used and valued across disciplines. As we navigate the evolving landscape of scholarly communication, such insights are crucial for optimizing the impact of data papers and fostering a more data-centric research ecosystem.

Methods

We took our sample from the journal of Data in Brief, an exclusively data journal founded by Elsevier in 2014. As one of the few journals that only publishes data papers, it has been frequently analysed to understand how research data is published and the relationship between data publication and scientific research (Chen et al., 2022; Fu et al., 2023; Li & Jiao, 2022; Thelwall, 2020).

We retrieved the metadata and citations of all data papers in the journal from the Dimensions database (https://www.dimensions.ai/). Dimensions is a well-established scholarly database that contains more than 146 million publications and have been heavily used in various scientometrics and science of science studies (Herzog et al., 2020; Martín-Martín et al., 2021). We only considered all data papers that have at least 10 citations indexed in the Dimensions database, so that we will have enough citations to analyse for each paper.

We specifically traced the domain information of all data papers from Dimensions that is based on the Australian and New Zealand Standard Research Classification (ANZSRC). It contains 22 broad divisions (FoR2) and 159 detailed groups of these divisions (FoR4). This classification was implemented in Dimensions on the paper-level using a machine learning algorithm (https://plus.dimensions.ai/support/solutions/articles/23000018820-which-research-categories-and-classification-schemes-are-available-in-dimensions), which can more accurately categorize the discipline of data papers published in the same journal than journal-level classification.

We mapped the FoR2 categories into five major knowledge domains for the next step of analysis. These five categories include: biomedical (*Bio*), environmental science (*Env*), physical science (*Phy*), social science and humanities (*Soc*), and Technologies and engineering (*Tech*). The mapping table, as well as the size of publications in each FoR2 category, can be accessed in our complementary materials from Zenodo (Li et al., 2024). For each of these five categories, we randomly selected 10 data papers into our final sample. And for each data paper, we randomly selected 10 publications that cite the paper to analyse in the next step.

We manually examined and filtered out the following research papers from the last step: (1) any publication that is not a research paper, such as review paper, data paper, and corrections, (2) any publication that is not published in English, and (3) any publication that is not citing the target data paper, despite the citation information supplied by the Dimensions database. Some of these issues are derived from the incomplete or inaccurate metadata information from the Dimensions database, which is a major issue in the research infrastructure to support research data. After this step, 309 research papers remain in our final sample.

For every publication, we manually collected sentences where an original data paper was cited. There are 437 sentences in total, given that a paper maybe cited multiple times in the citing publication. Two coders independently annotated every sentence using the following scheme we developed using another 50 randomly selected sentences taken prior to the final sample. The scheme focuses on distinguishing sentences describing or mentioning the data described in the data paper (i.e., data description, data use, and data background) or those that are not related to the data at all (i.e., research concept or background, research method, and research finding). Beyond the definitions presented in Table 1, examples of these categories can be accessed in our complementary materials shared on Zenodo (Li et al., 2024).

Category	Definition
Data-related	
Data Description (DP)	The sentence describes the dataset(s) per se, without indicating actual usage or how it was collected.
Data Use (DU)	The sentence describes the dataset is used in the research, including being used as a baseline.
Data Background (DB)	The sentence provides contextual information about the dataset, without describing its specific contents or usage.
Not data-related	
Research Concept or Background (NDC)	The sentence references the research topic or theoretical concepts of the cited data papers.
Research Method (NDM)	The sentence references the research methods, including but not limited to process, tools, techniques, used in the cited data paper.
Research Finding (NDF)	The sentence references the research findings, including results and their implications, derived from data in the cited data paper.

Table 1. Scheme for manual annotation of all citation sentences

We measured the inter-coder reliability of the annotation using Cohen's kappa implemented in the '*psych*' package of R (Revelle, 2024). The unweighted Cohen's kappa value is 0.73, which suggests substantial agreement (McHugh, 2012). We revisited sentences that are disagreed by the two coders.

Results

How are data papers cited in research articles, and for what purposes?

Our results show that the majority of data papers are NOT cited as research data in research articles. Table 2 shows the distribution of all citation sentences across the six categories. Among all 437 citations, only 48 of them (11.0%) belong to the data-related categories. And the largest category is *Research finding* (*NDF*), despite the fact that no data paper is supposed to present research design and findings based on its original definition (Chaven & Penev, 2011). However, there is also a decent share of citations to data papers that concerns research methods (i.e., *NDM*). Even though data papers are not cited as research data in these cases, this citation context is still relatively close to research data. Overall, we find a diverse spectrum of citation practice around data papers that are strongly deviant from the original purposes of publishing data papers.

Category	# Sentences	Share of Sentences
Data-related	**48**	**11.0%**
Data Use (DU)	27	6.2%
Data Description (DP)	14	3.2%
Data Background (DB)	7	1.6%
Not data-related	**425**	**89.0%**
Research Finding (NDF)	207	47.4%
Research Method (NDM)	130	29.7%
Research Concept or Background (NDC)	52	11.9%
All categories	**473**	**100%**

Table 2. Distribution of all citation sentences across the six categories

The above numbers on the level of sentences are translated into 16 data papers with at least one data-related citation. Among the 16 papers, only four of them are cited primarily data as research data (i.e., more than 50% of all citation sentences a paper received are data-related). Even though the reasons for such differences between data papers are not covered by this preliminary research, this would be an interesting question for the next step of our investigation.

We specifically examined how citation contexts are related to the paper sections in which a data paper is cited, given the strong relationship between these variables in citation analysis (Tahamtan & Bornmann, 2019). We classified all paper sections into the following four categories: Introduction (including literature review), Methods, RDC (i.e., Results, Discussion, and Conclusion), and Others (i.e., appendix and acknowledgement). The key statistics are presented in Table 3. For both data and method citations, they are more prevalent in the Methods section. In addition, even though only a few data papers are cited in appendix and acknowledgement, this section is also quite strongly connected to the data-related citation contexts.

Section	# Sentences	% Data citation	% Method citation
Introduction	149	3.4%	21.5%
Methods	106	23.6%	92.6%
RDC	158	4.4%	15.2%
Others	24	45.8%	7.7%

Table 3. Data citation index by paper section (% Data citation is the share of all data-related sentences among all sentences; % method citation is the share of all research meth-od sentences among all non-data sentences.)

Key factors behind the citation behavior: discipline and the citing-cited year gap

We examined two key factors (i.e., disciplines, citation year gaps) behind the citation contexts of data papers, which is reported in this section.

Discipline of data papers

We analysed the relationship between citation contexts and the discipline of data papers based on the ANZSRC classification system. Table 4 presents the results, which shows largely similar share of data citation to data papers from all domains. The higher share of data citations in the category of Soc (social sciences and humanities) is largely attributed to the single paper '*Residential electric vehicle charging datasets from apartment buildings*' (Sørensen et al., 2021), which has all citations related to data. And we believe this categorization is highly debatable, which is further highlighted in our discussion.

Discipline	# Sentences	% Data citation
Soc	75	14.7%
Phy	76	11.8%
Bio	72	11.1%
Tech	101	9.9%
Env	113	8.8%

Table 4. Data citation context by the discipline of data papers

Year gap of citing paper

We further investigated how the year gap between data papers and citing papers could be related to the citation contexts. Results, as shown in Table 5, illustrates that when the year gap widens, there is a growing possibility that the data paper is cited as research data. However, even after five years of the data paper, the share of data citation is still only at 20%. Our findings show similar patterns with citation contexts of method papers and method objects (Li, 2021; Small, 2018), where older research instruments are more likely to be cited as instruments, possibly because of the demonstrated validity.

Section	# Sentences	% Data citation
Year 0-1	119	5.0%
Year 2-3	145	15.2%
Year 4-5	121	9.1%
Year > 5	45	20.0%

Table 5. Data citation context by citation year gap between citing and cited documents

To further understand how the two variables are statistically correlated to the outcome, we constructed a logistic regression model to understand the correlation between the above variables as well as the publication year of the data papers as ***independent variables*** and whether a sentence is data-related or not as the ***categorical dependent variable***. The summary of the regression model is shown in Figure 1, with the y-axis showing how much each category contributes to the possibility of a sentence being a data-related citation. It shows that comparing to the domain of biomedical science (which is the baseline not shown in the graph), all other domains are at least equally likely having data-related citations, even though none of the other domains are significantly more likely so. Similarly, comparing to the year of 2016, data papers published in the years of 2017 and 2021 are significantly less likely to be cited as data, which is likely due to the smaller data points in these two years. We also did not find the year gap between citing and cited papers to be a significant factor for data citation.

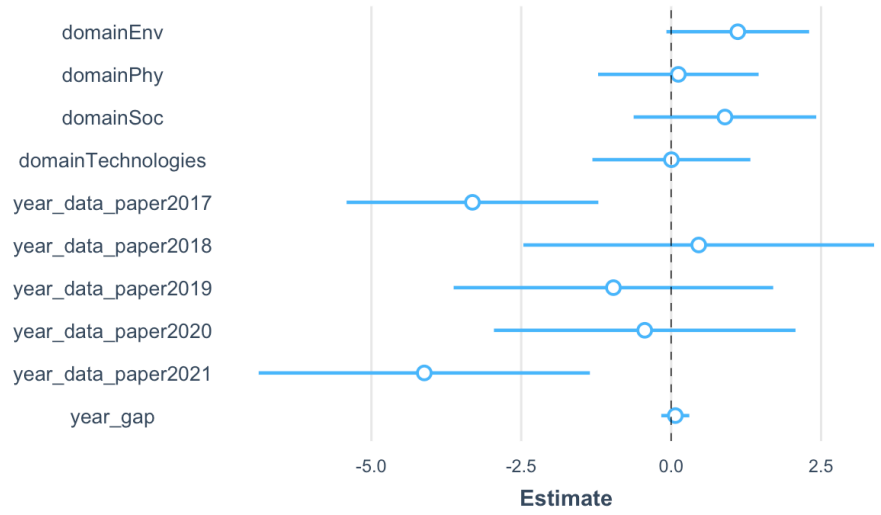


Figure 1. Summary of the statistical model

Discussion and conclusion

This short paper presents preliminary findings from our project to understand the citation contexts of data papers in English-language research articles. We examined the general patterns of whether data papers are cited as research data, an important argument to support the usefulness of this new genre, with special focus on some key factors behind the citation behavior, such as the discipline of the data paper and citation time gap. Our results reconfirm existing findings that data papers can be cited for various reasons beyond just those related to research data (Jiao & Darch, 2020; Thelwall, 2020). However, the fact that the majority of data papers are not cited as research data raises a critical concern about whether data papers can help to give the proper credit (i.e., credit of creating and publishing research data) to the authors.

Additionally, building upon existing efforts, our paper strives to offer a more comprehensive empirical investigation on factors behind the data paper citation behavior. We find that the general pattern stays valid across knowledge domains and does not change significantly as the year gap between the citing and cited documents widens.

Beyond our major findings, our research design and process also reveal some key issues in the infrastructure to support data publication. One critical issue is the identification of the discipline of data papers. While many data papers are published in multidisciplinary data journals, such as *Data in Brief* and *Scientific Data*, the journal-level classification system in many scholarly databases (such as the Web of Science) is inadequate for determining the discipline of such publications. Even by using a novel paper-level classification, we find cases where the classification may not be accurate, which could have strong impact on future quantitative studies on this topic. We argue that establishing a more robust and accurate system to evaluate the discipline of data papers and research data is a critical next step; and this system should consider not just paper metadata, but also the authors' affiliations and other attributes of the research datasets to achieve better performance.

In the next step of the project, we will expand the presented investigation by using a larger sample size as well as considering more factors to explain the citation behavior of data papers, such as the discipline of the citing paper and whether the citation is from the data paper authors themselves (i.e., self-citation). We expect that these extra factors will have strong impacts on how a data paper is understood and discussed in citing publications. In addition, we will also use advanced statistical

models to understand more factors behind the citation behavior of data papers, which will contribute to the literature of citation context analysis.

About the authors

Kai Li is Assistant Professor in School of Information Sciences, University of Tennessee, Knoxville, USA. He received his Ph.D. from the Department of Information Science at Drexel University, and his research interest are scholarly communication, quantitative science studies. He can be contacted at kli16@utk.edu.

Pao-Pei Huang is Ph.D. student in School of Information and Library Science, University of North Carolina, Chapel Hill, USA. Her research interest are open science practices, scholarly communication, and social informatics. She can be contacted at paopei@unc.edu.

Wei Jeng is Associate Professor at Department of Library and Information Science, National Taiwan University, Taiwan and Director of Talent Empowerment Center at National Institute of Cyber Security, Taiwan. She received her Ph.D. from the school of computing and information at University of Pittsburgh, and her research interest are open science practices and research data infrastructure. She can be contacted at wjeng@ntu.edu.tw.

References

Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, 12, 1-12. https://doi.org/10.1186/1471-2105-12-S15-S2

Chen, P. Y., Li, K., & Jiao, C. (2022). A preliminary analysis of geography of collaboration in data papers by S&T capacity index. *Proceedings of the Association for Information Science and Technology*, 59(1), 642-644. https://doi.org/10.1002/pra2.676

Fu, J., Tian, L., Zhang, C., & Li, J. (2023). Opening research data contributes to the citations of related research articles: Evidence from Data in Brief. *Learned Publishing*, 36(3), 426-438. https://doi.org/10.1002/leap.1551

Gorgolewski, K.J., Margulies, D.S., & Milham, M.P. (2013). Making data sharing count: A publication-based solution. *Frontiers in neuroscience*, 7, 9. https://doi.org/10.3389/fnins.2013.00009

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative science studies*, 1(1), 387-395. https://doi.org/10.1162/qss_a_00020

Jiao, C., & Darch, P.T. (2020). The role of the data paper in scholarly communication. *Proceedings of the Association for Information Science and Technology*, 57(1), p.e316. https://doi.org/10.1002/pra2.316

Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies. *F1000Research*, 3. https://doi.org/10.12688/f1000research.3979.3

Li, K. (2021). The re-instrumentalization of the Diagnostic and Statistical Manual of Mental Disorders (DSM) in psychological publications: A citation context analysis. *Quantitative Science Studies*, 2(2), 678-697. https://doi.org/10.1162/qss_a_00124

Li, K., & Jiao, C. (2022). The data paper as a sociolinguistic epistemic object: A con-tent analysis on the rhetorical moves used in data paper abstracts. *Journal of the As-sociation for Information Science and Technology*, 73(6), 834-846. https://doi.org/10.1002/asi.24585

Li, K., Huang, P. P., & Jeng, W. (2024) Dataset for "Are data papers cited as research data? Preliminary analysis on interdisciplinary data paper citations" [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13763303

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871-906. https://doi.org/10.1007/s11192-020-03690-4

McHugh, M.L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Revelle, W. (2024). Package 'psych' version 2.4.3. The comprehensive R archive net-work. https://CRAN.R-project.org/package=psych

Schöpfel, J., Farace, D., Prost, H., & Zane, A. (2020). Data papers as a new form of knowledge organization in the field of research data. *Knowledge Organization*, 46(8), 622-638. https://doi.org/10.5771/0943-7444-2019-8-622

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, 12(2), 461-480. https://doi.org/10.1016/j.joi.2018.03.007

Sørensen, Å.L., Lindberg, K.B., Sartori, I., & Andresen, I. (2021). Residential electric vehicle charging datasets from apartment buildings. *Data in Brief*, *36*, 107-105. https://doi.org/10.1016/j.dib.2021.107105

Stuart, D. (2017). Data bibliometrics: Metrics before norms. *Online Information Review*, *41*(3), 428-435. https://doi.org/10.1108/OIR-01-2017-0008

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*(3), 1635-1684. https://doi.org/10.1007/s11192-019-03243-4

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PloS one*, 6(6), p.e21101. https://doi.org/10.1371/journal.pone.0021101

Thelwall, M. (2020). Data in Brief: Can a mega-journal for data be useful?. *Scientometrics*, *124*(1), 697-709. https://doi.org/10.1007/s11192-020-03437-1

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

'A Mulligan's stew': educational preparation for today's academic library liaisons in the humanities and social sciences

Rachel Fleming-May, Wade Bishop, and Caroline Villarreal

DOI: https://doi.org/10.47989/ir30iConf46912

Abstract

Introduction. This short paper reports on findings from a survey of academic librarians with liaison responsibilities to college and university social sciences and humanities departments, with particular attention to newer responsibilities related to research data management and scholarly communications, areas less commonly associated with humanities and humanistic social sciences liaison work.

Method. The survey was distributed to 1085 individuals, 330 of whom responded (rate: 30.4%).

Analysis. This paper reports on one segment of the survey's findings: the extent to which respondents believe their American Library Association (ALA)-accredited master's degree and, where applicable, additional graduate coursework, prepared them for specific duties associated with liaison work.

Results. Three findings are discussed: 1) impressions about the ALA-accredited master's from respondents who have not participated in other graduate education 2) as compared to respondents who have completed additional graduate education, and 3) respondents' impressions of the extent to which additional graduate study prepared them for liaison work as compared to the preparation afforded by the ALA-accredited master's.

Conclusions. Respondents do not believe the ALA-accredited master's to have provided strong preparation for the '*Mulligan's Stew*' of tasks, skills, and responsibilities associated with liaison work. The Discussion and Conclusion sections share suggestions for strengthening liaison preparation.

Introduction

Liaison to academic departments remains one of the most common functions of today's academic libraries (Johnson, 2018). Also referred to as '*subject specialist*, *subject librarian*, *academic liaison*, *subject liaison*, *liaison librarian*, and *subject bibliographer*' liaison librarians have traditionally assisted students, faculty, and staff in assigned academic departments with materials, research, and instructional needs (Church-Duran, 2017, p. 258). In addition to providing these critical services to their constituents, liaisons constitute the public face of the library, offering opportunities to publicize library facilities, resources, and services. However, faculty and students' needs have continuously expanded and changed to include support for data management, copyright, document repository, and open educational resources (OER's). Academic libraries have enlisted a variety of strategies to support these requirements, but the extent to which library and information science (LIS) master's-level programs have evolved to support the expanded educational needs of students preparing for careers as academic library liaisons is unclear.

As educating future library professionals remain a core function of American Library Association (ALA)-accredited master's degree programs, there is sustained interest in exploring the extent to which these programs are preparing students for employment in today's libraries. Recent research has explored library professionals' impressions of the preparation they received for work in areas including technical services (Smith, et al., 2024), data management (Luo and Tang, 2024; Thomas and Urban, 2018), liaison to academic departments (Bright and Colón-Aguirre, 2022), and special collections and rare book cataloguing (Hertenstein, 2023). Findings from such studies provide useful guidance for continued evolution of LIS education as they sketch a realistic picture of the skills and knowledge required for specific specializations.

This paper reports on findings from a survey of academic librarians with liaison responsibilities to college and university humanities departments. The survey included questions about both typical and non-traditional liaison responsibilities. Respondents who indicated responsibility for specific duties were also asked to evaluate the preparation for those tasks afforded them in both their ALA-accredited master's and, where applicable, other graduate degree programs (a small number of respondents indicated not having enrolled in an ALA-accredited master's or being in the process of earning the degree). This area of questioning is this paper's focus, and therefore reports on the sub-set of survey respondents who have completed an ALA-accredited master's degree and indicated responsibility for the tasks discussed here.

Literature review

The changing liaison role

The role of liaison has traditionally been assigned to librarians with subject expertise or interest in specific academic departments outside the library. Traditionally, liaison responsibilities have '*included outreach*, *collection development*, *instruction*, *scholarly communication support*, *and reference*' (Bright and Colón-Aguirre, 2022, p. 2). As user needs have changed, academic libraries have developed new services, such as providing scholarly impact and metrics information, promoting the institutional repository, and research data management. These new responsibilities are not always assigned to liaison librarians, however: many academic libraries have chosen to distribute responsibility for newer areas of academic library service to '*functional specialists*,' individuals whose '*expertise is not in subject content but rather in tools, methods, and practical domains…*' (Narlock & Robinson, 2022, p. 81). Functional specialists are meant to address these needs on an institution-wide, discipline-agnostic basis. For example, while a department's liaison librarian would be expected to assist faculty members with specialized research queries, they also might enlist the assistance of their library's data librarian for help with any data-specific needs that arise. In the case of research data management, requests for assistance with locating, assessing, cleaning, storing, and meeting standards for data have grown in recent years, leaving small

functional specialist departments overwhelmed with requests for assistance. In addition to the difficulty presented by the sheer number of requests, it can be difficult for a functional specialist to develop and maintain knowledge of multiple sets of requirements, conventions, and customs across scores of disciplines.

Some libraries are transforming traditional liaison work to meet constituents' needs, primarily by reshaping the traditional liaison model in favor of a task-based team approach. Flaxbart (2018) provided a case study of such a reconfiguration at the University of Texas, which reorganized liaisons into three teams: 1) arts, humanities, and global studies (eight liaisons) and 2) STEM and social sciences (seven liaisons), both in the research support and digital initiatives department, and 3) teaching and learning in the teaching and learning services department (six liaisons). Flaxbart adds four liaisons were reassigned to the scholarly resources department to manage core collections in arts and humanities, global studies, STEM, and social sciences, respectively, presumably removing the traditionally-core collection management responsibility from the 19 liaisons reassigned to the three teams (p. 73). Frenkel, et al. (2018) describe a similar process at the University of California-Riverside which saw subject specialist librarians reorganized into three departments: Teaching and Learning, tasked with providing instruction and services to support instruction, Research Services, to '*provide support across the research life cycle*' (p. 53), and Collection Strategies, tasked with developing '*data-driven collection management strategies and communication methods*' to support the university's core functions (p. 57). In these new models, liaisons have responsibility for data management, digital initiatives, and supporting new efforts at measuring research impact. Services provided by library liaisons are both reactive and proactive. In the case of the former, librarians respond to requests from users, while in the latter, librarians offer users services that they might not be aware they would find useful, or often, that they might not understand what the library and their liaisons can provide. This may be true of humanists with research data management needs, as '*few humanists think of their sources as data*' (Sanders, 2024, p. 101).

Optimal education for the liaison role

Expanding the services expected of liaisons begs the question: what preparation have the individuals tasked with these new responsibilities received? Specifically, did the ALA-accredited master's afford adequate qualifications for success in these evolving areas of service? Bright and Colón-Aguirre (2022) note that educational programs such as Collaborative Analysis Liaison Librarians (CALL) have emerged to prepare students for a more inclusive view of liaison work, but CALL was designed for science liaisons. What about liaisons to disciplines in the social sciences or humanities? While requirements for individual institutions and positions vary, the ALA-accredited master's degree remains the baseline requirement for librarians employed in most academic libraries. The LIS literature is rife with practitioners' reflections on the preparation for employment afforded by their formal graduate education. In response to the statement '*[m]y LIS program's curriculum prepared me to be a liaison librarian,*' only 25.8% and 7.1% of Bright and Colón-Aguirre's (2022) survey respondents (respectively) agreed or agreed strongly. Bright and Colón-Aguirre (2022) used a survey and follow-up interviews to gather feedback from liaison librarians about the preparation afforded them by that degree. Those librarians who completed a curricular '*track*' designed for academic librarianship were more likely to believe for-credit courses provided some of the knowledge necessary to be successful as a liaison, particularly through those aligned with '*traditional*' liaison work: '*instruction, reference, collection development, and specialized courses.*' (p. 7) Respondents were significantly more likely, however, to point to co-curricular activities—assistantships, internships, and part time work—as contributing more to their readiness for the liaison role than did their formal coursework.

Explorations of the most desirable education for liaison librarians is not limited to the ALA-accredited master's: despite the Association of College and Research Libraries' 1975 assertion

(reaffirmed in 2001, 2007, 2011, and 2018) that the ALA-accredited master's is the appropriate terminal degree for academic librarians (Association of College and Research Libraries, 2018), discussions of the desirability or necessity of additional graduate education has featured heavily in the LIS literature, often pointing to benefits an additional graduate degree confers to an academic librarian, including improved acculturation to academic life, strengthened subject knowledge, and increased visibility among and acceptance by faculty members (Fleming-May and Douglass, 2014, p. 396). Several also identify improved understanding of and preparation for the process of conducting research as a significant boon of graduate study beyond the ALA-accredited master's (Gilman and Lindquist, 2010, p. 403; Mayer and Terrill, 2005, p. 64). In their study comparing opinions of librarians with a second graduate degree to those with only the ALA-accredited master's, Day and Szurek (2018) concluded that the value of a second graduate degree for subject liaisons is '*debatable*'. (p. 142) Although both groups of respondents reported engaging with liaison duties at similar rates, a large majority of respondents with the second graduate degree indicated believing it enhanced their ability to liaise (87%) and provide instruction (74%) (p. 139). On the other hand, only 44% of those without a second graduate degree believed it would be beneficial for liaison work (p. 144). Sevryugina, et al. (2024) surveyed a group of academic librarians, some of whom also held a subject Ph.D. Nearly half the survey respondents were employed as subject liaisons. Responses to their survey's question about the usefulness of a Ph.D. were similar to Day and Szurek's (2018) findings: while only 20% of non-Ph.D. holders answered that a doctoral degree would be useful to librarians, 67% of Ph.D.-holding librarians answered that question in the affirmative (p. 370). Echoing previous studies' findings, Sevryugina, et al.'s respondents provided similar examples of areas of the academic liaison librarian's repertoire for which a subject graduate degree would be helpful, such as improved disciplinary expertise, greater experience/facility with research, and pedagogical knowledge. (p. 371). These studies of education for liaison work do not typically discuss the '*functional specializations*,' that this paper is concerned with.

Method

This paper builds upon the tradition of research into librarians' opinions about the extent to which graduate education prepared them for their work by reporting on a survey of academic library liaisons to humanities and social sciences academic departments. Notably, the survey is concerned primarily with liaisons' preparation for several areas of librarianship typically assigned to functional specialists, described by one respondent as a '*Mulligan's stew*.' These can be grouped under the broad categories of data management, assessment, and scholarly communications. Wishing to test our hypothesis that librarians assigned as liaisons to humanities and social sciences departments might not have received academic preparation in these areas, we distributed survey invitations to librarians who 1) are listed as humanities department liaisons on the web directories of Association of Research Libraries (ARL) member institutions in the United States, and/or 2) are listed as humanities department liaisons on the web directories of colleges and universities identified by the National Endowment for the Humanities (NEH) as home institutions of recipients of one or more NEH Office of digital humanities awards made between 2011 and 2024.

Invitation	Number Sent	Responses	Rate
NEH-Awarded Institutions	696	217	31.2%
ARL Member Institutions (not home to NEH Awarded programs)	389	113	29%

Table 1. Distribution of invitations and survey response rate

This paper focuses on one dimension of the survey: questions about respondents' perceptions of the extent to which their graduate education prepared them for several areas of liaison work as identified in previous surveys of liaison librarians (Bishop, et al., 2023) and working documents, such as the University of North Carolina, Greensboro Libraries' "Liaison Roles and Responsibilities"

(UNCG Libraries, undated). In response to the question, '*To what extent did your ALA-accredited master's program prepare you to…*' respondents were asked to select '*well*', '*somewhat*', or '*did not*' for each item:

1. Understand the role of library liaison in higher education
2. Understand the specific research processes of the academic disciplines you work with
3. Keep current with general trends in scholarly communications
4. Monitor subject-specific trends in scholarly communications
5. Communicate with researchers (faculty and graduate students) regarding their research support needs
6. Contribute to accreditation reports and applications for new academic programs
7. Support donor connections as relevant to liaison subject areas
8. Inform constituents about scholarly communication issues: copyright, author rights, etc.
9. Encourage and support the writing of data management plans
10. Inform constituents about issues related to open access publishing, institutional repositories, etc.
11. Analyze user data (i.e., reference chat logs, COUNTER usage reports) for trends
12. Monitor research and publishing trends in specific subject areas
13. Examine papers authored by constituents in specific subject areas for research interests, trends, and use of research sources

Findings

Responses to this group of questions both echoed and deviated from other studies' findings discussed in the literature review. Figure 1 represents survey responses divided into two groups: those from liaison librarians holding only the ALA-accredited master's ('ALA Only'), and those holding an additional graduate degree ('ALA +'). The numbers in parentheses in the chart refer to the number of responses to each item. Respondents could select "not applicable" for those tasks for which they were not responsible.

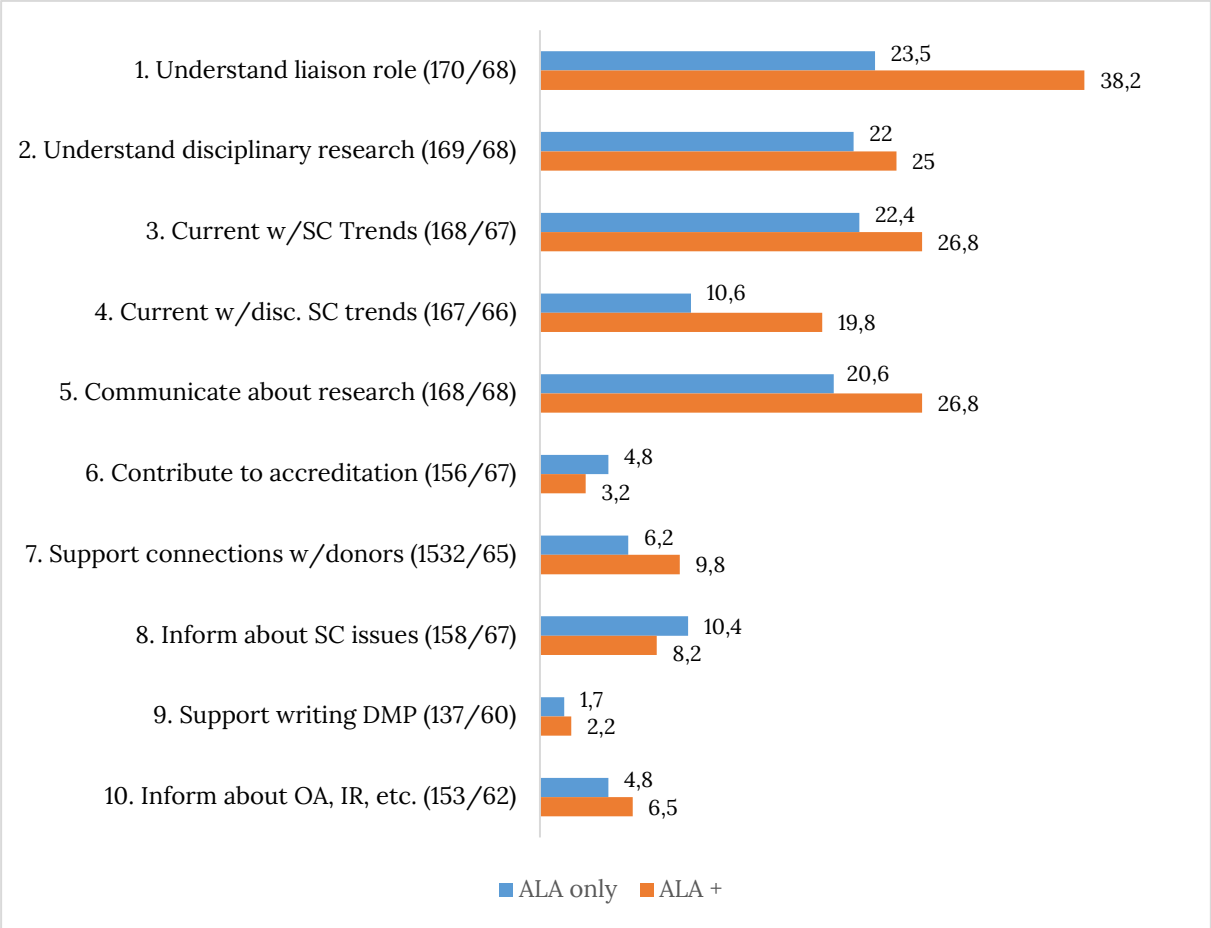


Figure 1. 'My ALA-accredited master's degree prepared me well.'

Figure 1 depicts respondents' impressions of the quality of preparation they received for specific liaison tasks during the ALA-accredited masters. Indeed, in the case of the item for which the largest percentage of respondents indicated the ALA-accredited master's prepared them *'well'* '*understand[ing] the role of library liaison in higher education*', this option was selected by only 38.2% of those respondents with an additional graduate degree, and only 23.5% of those with the ALA-accredited master's. This item does illustrate an interesting pattern in survey results, however: those respondents who have completed other graduate work in addition to the ALA-accredited master's seemed to view that degree in a more favourable light than those who had only completed the ALA-accredited master's. The percentage of dual-degree holders who categorized themselves as well-prepared by the ALA-accredited master's to '*keep current with disciplinary trends in scholarly communication*,' '*keep current with general trends in scholarly communication*', and '*understand the role of library liaison in higher education*' are all higher than their single-degreed counterparts. We can only speculate about the reason for this discrepancy: perhaps librarians who have completed graduate work outside the ALA-accredited degree have more realistic expectations for graduate education in general?

We also asked those respondents with an additional graduate degree to assess how well that program prepared them for the same tasks. Figure 2 compares the percentage of respondents who selected '*well*' to fill in the statement '*my ALA accredited master's degree/other graduate degree prepared me…*' for each of the tasks. Note: '*ALA degree*' category includes both groups of respondents in figure 1. Respondents were invited to select 'N/A' if they were not responsible for

a specific task among their liaison responsibilities, so the number of responses for individual items vary; these are indicated in the numbers in parentheses.

Note: 'ALA *Degree*' category includes both groups of respondents in figure 1. In Figure 2, responses from the two groups represented in Figure 1 have been collapsed. Figure 2 also includes three items that did not appear in the segment of the survey administered to those respondents who indicated not having completed additional graduate work: '*examin[ing] papers authored by constituents in specific subject areas for research interests, trends, and use of research sources*,' '*analys[ing] user data (i.e., reference chat logs, COUNTER usage reports) for trends*,' and '*monitor[ing] research and publishing trends in specific subject areas*.' Consequently, those items represent a smaller number of respondents.

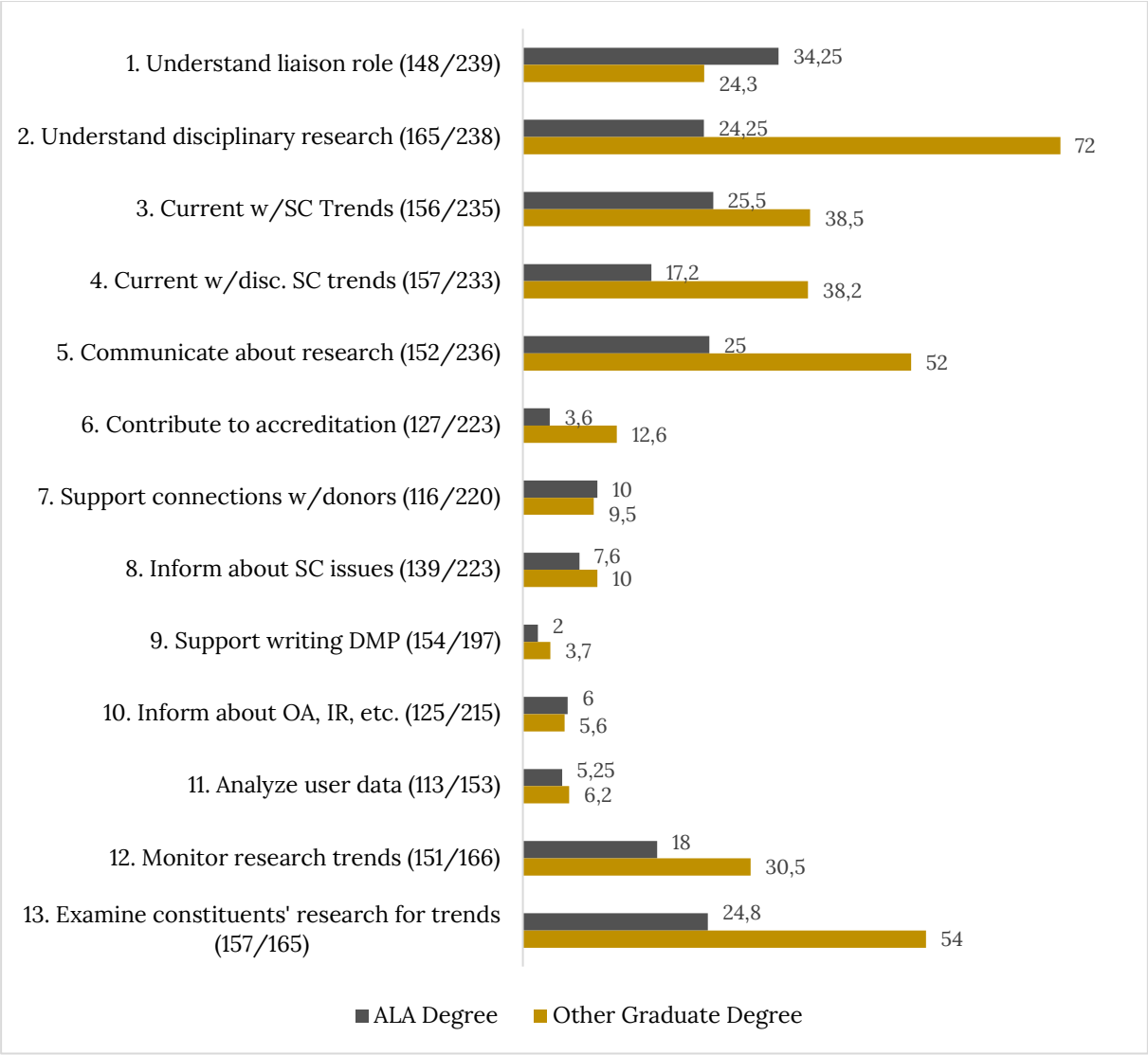


Figure 2. 'My other graduate/ALA-accredited master's degree prepared me well to…'

Comparing respondents' assessment of the preparation afforded by the ALA-accredited master's to that provided by other graduate education is eye-opening as well. Only in the case of the item '*understand the role of library liaison in higher education*' did a notably larger percentage of respondents indicate their ALA-accredited master's program prepared them '*well*' than did their subject graduate work. Respondents considered themselves well prepared by other graduate work

in several areas, including '*understand[ing] the specific research processes of the academic disciplines you work with*', '*communicat[ing] with researchers (faculty and graduate students) regarding their research support needs*', and '*examin[ing] papers authored by constituents in specific subject areas for research interests, trends, and use of research sources.*'

It is interesting to note, however, that respondents did not consider themselves well-prepared by either graduate program in several areas, primarily those typically considered to be under the purview of functional specialists: '*analys[ing] user data for trends*' (electronic resources management), '*inform[ing] constituents about issues related to open access publishing, institutional repositories, etc.*' (scholarly communications), '*encourag[ing] and support[ing] the writing of data management plans*' (research data management), and '*inform[ing] constituents about scholarly communication issues: copyright, author rights, etc.*' (scholarly communications/copyright). The item '*support[ing] donor connections as relevant to liaison subject areas*' could be seen as the responsibility of those library staff tasked with fundraising or soliciting materials donations for special collections. Liaisons' subject knowledge and familiarity with their counterparts in academic departments, however, could make them strong contributors to these efforts.

Conclusions

Responses to our survey suggest LIS programs could improve guidance for students' interest in becoming academic library liaisons. Increasing curricular offerings where gaps between additional graduate degree holders and those who have only earned the ALA-accredited master's (e.g., research awareness and understanding) could be particularly beneficial. It appears that LIS education and graduate education in other disciplines have been equally remiss in educating students in functional specializations that could be '*new liaison*' areas such as data management and scholarly communications. Here, LIS education has an opportunity to stake a position ourselves to our institutional colleagues as expert educators and consultants. Responses to this area of inquiry provide fascinating breadcrumbs for both LIS educators seeking to improve their programs' preparation for librarians entering academic librarianship and for librarians and library administrators interested in identifying professional development opportunities.

Acknowledgements

The authors would like to thank the University of Tennessee, Knoxville for its support. We are grateful for the referees' thorough attention to our paper, and for their helpful comments and suggestions for improvement. We would also like to acknowledge our appreciation for our study's respondents' frank and thoughtful survey responses.

About the authors

Rachel Fleming-May is an Associate Professor in the School of Information Sciences at the University of Tennessee, Knoxville. Her research interests include the evolving nature of academic librarianship. She can be contacted at rfm@utk.edu

Wade Bishop is a Professor in the School of Information Sciences at the University of Tennessee, Knoxville. His research foci include research data management, data discovery, geographic information science, as well as the study of data occupations, education, and training. He has published many works evaluating the services and resources of academic and public libraries. He can be contacted at bbisho13@utk.edu

Caroline Villarreal is an Adult Services Librarian at the Berkshire Athenaeum in Pittsfield, Massachusetts. She received her Master of Science in Information Sciences from the University of Tennessee, Knoxville. She can be contacted at caroline@pittsfieldlibrary.org.

References

Association of College and Research Libraries. (2018). *Statement on the Terminal Professional Degree for Academic Librarians*. https://www.ala.org/acrl/standards/statementterminal

Bishop, B. W., Gavel, S., Chapin, E., Fernandez, P., & Allard, S. (2023). Science, technology, engineering, and mathematics (STEM) liaison librarians: Perspectives on functions and frequencies for serving academic researchers. *Library & Information Science Research*, 45(4), 101265. https://doi.org/10.1016/j.lisr.2023.101265

Bright, K. M., & Colón-Aguirre, M. (2022). Prepare to be unprepared? LIS curriculum and academic liaison preparation. *The Journal of Academic Librarianship*, 48(6), 102602. https://doi.org/10.1016/j.acalib.2022.102602

Church-Duran, J. (2017). Distinctive Roles: Engagement, Innovation, and the Liaison Model. *Portal: Libraries and the Academy*, 17(2), 257–271.

Day, L., & Szurek, J. (2018). Beyond the MLS: Reflections on the role of professional and continued education for academic library liaisons. *College & Undergraduate Libraries*, 25(2), 131–149. https://doi.org/10.1080/10691316.2018.1455547

Flaxbart, J. (2018). Iterative Evolution of the Liaison Librarian Role: A Brief Case Study. *Research Library Issues*, 294, 72–75. https://doi.org/10.29242/rli.294.6

Fleming-May, R. A., & Douglass, K. (2014). Framing Librarianship in the Academy: An Analysis Using Bolman and Deal's Model of Organizations. *College & Research Libraries*, 75(3), forthcoming.

Frenkel, A., Moxham, T., Cook, D. B., & Marshall, B. (2018). Moving from Subject Specialists to a Functional Model. *Research Library Issues*, 294, 39–71. https://doi.org/10.29242/rli.294.5

Gilman, T., & Lindquist, T. (2010). Academic/Research Librarians with Subject Doctorates: Experiences and Perceptions, 1965-2006. *Portal: Libraries and the Academy*, 10(4), 399–412.

Hertenstein, L. (2023). Current State of Special Collections and Rare Books Cataloging Education at LIS Programs. *Cataloging & Classification Quarterly*, 61(7–8), 773–791. https://doi.org/10.1080/01639374.2023.2247402

Johnson, A. M. (2019). Connections, Conversations, and Visibility: How the Work of Academic Reference and Liaison Librarians Is Evolving. *Reference & User Services Quarterly*, 58(2), Article 2. https://doi.org/10.5860/rusq.58.2.6929

Luo, J., & Tang, R. (2024). Data competency for academic librarians: Evaluating present trends and future prospects. *The Journal of Academic Librarianship*, 50(4), 102897. https://doi.org/10.1016/j.acalib.2024.102897

Mayer, J., & Terrill, L. J. (2005). Academic Librarians' Attitudes about Advanced-Subject Degrees. *College & Research Libraries*, 66(1), Article 1. https://doi.org/10.5860/crl.66.1.59

Narlock, M., & Robison, M. (2022). Liaison Librarianship in Shiny Packages: An Exploration of Product Ownership in Academic Libraries. *The International Information & Library Review*, 54(1), 80–87. https://doi.org/10.1080/10572317.2022.2025672

Sanders, A. R. (2024). Humanistic Data—Classifying Individuals and Visualizing Silences. In A. R. Sanders (Ed.), *Visualizing History's Fragments: A Computational Approach to Humanistic Research* (pp. 101–144). Springer International Publishing. https://doi.org/10.1007/978-3-031-46976-3_4

Sevryugina, Y. V., Smith, C., & Mordoch, G. (2024). Challenges, Successes, and Possibilities: How MLS- and PhD-holding Librarians Perceive their Roles in Academic Libraries. *Portal: Libraries and the Academy*, 24(2), 361–403.

Smith, C., Daugherty, A. L., & Lowry, L. (2024). Technically prepared: Librarians' perceptions on LIS curricula and technical services workforce preparedness. *The Journal of Academic Librarianship*, 50(1), 102815. https://doi.org/10.1016/j.acalib.2023.102815

Thomas, C. V. L., & Urban, R. J. (2018). What Do Data Librarians Think of the MLIS? Professionals' Perceptions of Knowledge Transfer, Trends, and Challenges. *College & Research Libraries*, 79(3), Article 3. https://doi.org/10.5860/crl.79.3.401*Liaison Roles and Responsibilities*. (n.d.).

University of North Carolina, Greensboro University Libraries. Liaison Roles and Responsibilities. Retrieved January 7, 2025, from https://library.uncg.edu/about-us/libraries-liaison-program/liaison-roles-and-responsibilities/

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

AI literature review systems: an analysis of performance, affordances, and outputs for a complex topic in the social sciences

Heather Moulaison-Sandy, Wilson Castaño-Muñoz, Laura Ridenour, and Denice Adkins

DOI: https://doi.org/10.47989/ir30iConf46906

Abstract

Introduction. Artificial intelligence has the potential to revolutionize the way that scholars produce work, including through the creation of literature reviews using free or low-cost systems. Despite the importance of the literature review, AI-powered tools to create literature reviews are understudied at present.

Method. To address this gap, four specialized tools were assessed. The complex area of study chosen was e-reading in English and Spanish, a field where differences are evident in the two languages. Systems were prompted to write a literature review of 500-1000 words, and the resulting outputs were analysed based on the requirements for reviews of the literature.

Results. Although none of the systems allowed for all the identified criteria to be addressed, Scite allowed many criteria to be met and was able to produce text in Spanish as well as in English. The remaining three systems investigated, Jenni.ai, OpenRead, and Wisio, performed less favourably.

Conclusions. AI-powered systems show incredible promise, with one unstudied area being the creation of scholarly reviews of the literature. Although the kinds of research questions that might be examined in social sciences like LIS will necessarily be complex, until now little has been done to evaluate these systems.

Introduction

Artificial intelligence (AI) systems show promise in supporting scholars in the discovery and analysis of articles that are part of the scholarly record. A suite of AI-powered tools is emerging to support the creation of reviews of the literature. Specifically, these systems advertise the ability to analyse articles that will create new meaning, identify lacuna in the scholarly record, and support the understanding of the state of the field and, by extension, further steps that should be taken. By including scholarly sources, these systems allow researchers to explore the literature in their field, often for free or very low cost.

The performance, outputs, and affordances of these systems has only been evaluated anecdotally. Reddit threads have probed this question, in some cases with limited success (e.g., ijmarkandel, 2023). In other words, although the accuracy of aspects of certain systems (e.g., references in ChatGPT (Mugaanyi et al., 2024)) have been evaluated, very little is known about free and low-cost AI systems that support the creation of complex reviews of the scholarly literature, and no formal studies have been conducted.

In the social sciences, complexities readily emerge when studying the scholarly record relating to the position of the authors. Different terms can be used to describe a single phenomenon and different methodologies and literatures can be sources; language traditions also have different approaches to scholarship. Anecdotally, electronic reading (also known as e-reading, digital reading, or reading on screen) is one topic that is studied differently in anglophone and Hispanophone traditions. It represents a complex, social sciences-based problem that has not adequately been addressed in the literature.

This paper focuses on the novel topic of the performance of AI systems and their functionalities relative to traditional work in creating reviews of the literature in the social sciences, especially when working in novel, complex spaces. The complex topic explored is e-reading as studied in the scholarly literature in English and Spanish.

Area for application: e-reading

Some specific concepts in the field of digital reading have been covered using different terms. One such concept is the process by which a new medium reshapes an established one, simultaneously referencing and reinterpreting its conventions and traditions, termed remediation in English (Bolter & Grusin, 1996). Works in Spanish also address this concept in relation to the book and its evolution to the digital book, focusing on the materiality, format and features of the new medium. However, studies in Spanish do not use the term *remediación*, nor do they cite the English-language studies that coined the term (e.g., Canclini et al., 2015; Cavallo & Chartier, 1999; Cordón, 2016; Morales & Espinoza, 2003).

Given that with e-reading, researchers from different language families address similar concepts using different terminology and by citing different sources, understudied complexities emerge. To address this, we investigate the ability for free and low-cost AI-powered systems to create complex reviews of the literature. To our knowledge, although the topic of AI-powered tools and systems have been analysed for their usability in academia, no scholarly papers investigate tools for reviews of the literature by analysing their performance and affordances. The current project aims to address this gap in the literature.

Research question

To investigate this topic, the following research question is explored:

> RQ1: To what extent are AI systems able to carry out sophisticated reviews of the literature for use in the social sciences?

Review of the literature

Literature reviews in the social sciences

In scholarship, reviews of the literature ground a project in the extant literature, situating it and allowing it to contribute to the field (Machi, L. A. & McEvoy, 2016). High quality literature reviews organize and synthesise completed work (Zhao et al., 2024). They help to identify a '*gap*' and ideally are shaped to provide an argument as to what scholars in a given field should be able to do in the future (Slater, 2018). Practically, the sources will be carefully chosen, and elements including source, author, and date will play a role. The perspective adopted will support the research questions and will limit the scope of the work meaningfully, presenting an argument (Machi & McEvoy, 2016); in making choices to include sources, the extent of the analysis of the results and the scope of the discussion are likewise limited. Developing the ability to craft such an argument is a skill that takes years for a human researcher or scholar to develop (Joos et al., 2024; Slater, 2018).

AI-powered systems to support scholarship

AI-powered tools have emerged that support scholarship. This paper focuses on a set of AI-powered tools that allows for the automated creation of literature reviews or other kinds of analyses of scholarly articles. Because literature reviews are related to the discovery of articles and the connectedness of articles, these tools are related to and often overlap with other sets of AI-powered tools.

For example, to carry out a literature review, scholars must first identify relevant scholarly articles. One example of a searchable corpus is SemanticScholar (https://www.semanticscholar.org). The AI-powered search engine Consensus (https://consensus.app/) makes use of this corpus (Consensus AI, 2022). Scopus AI (https://www.elsevier.com/products/scopus/scopus-ai) has a similar approach. Other AI-powered search engines exist (e.g., Supersymmetry (https://www.supersymmetry.ai/)) that also support enhanced search and discovery. Lateral (https://www.lateral.io/) and Scispace (https://typeset.io/) support discovery and compiling notes.

Researchers can benefit from other support as well, and many systems have multiple ways of supporting authors. Another class of AI-powered tools is systems designed to help with the analysis and summarization of articles. For example, ChatPDF (https://www.chatpdf.com/), Explainpaper (https://www.explainpaper.com/), Humanata (https://www.humata.ai/), ResearchMate Pro (https://researchmate.pro/), Scholarcy (https://www.scholarcy.com/), Sharly (https://sharly.ai/), StudyFlo (https://www.studyflo.com/), and Unriddle (https://www.unriddle.ai) support quick overviews of articles, in some cases if the publisher is willing to expose the source to analysis. Other AI-powered systems enable the discovery or visualization of connections between articles, including Connected Papers (https://www.connectedpapers.com/), Inciteful (https://inciteful.xyz/), Litmaps (https://www.litmaps.com/), and ResearchRabbit (https://www.researchrabbit.ai/).

Overview: AI-Powered systems supporting scholars and literature-based discovery

Artificial Intelligence Systems used for developing literature reviews are largely some forms of Large Language Model (Shopovski, 2024). One of the basic tenets of scholarly communication is the need for scholars to position their work in the field of endeavor. Goals of using LLMs is to streamline the process of surveying the literature (Joos et al., 2024), identify documents for inclusion in a review (Antu et al., 2023), evaluate and categorize relevant documents (Joos et al., 2024), interpret documents in the context of research, code relevant documents (Joos et al., 2024), and aid in the human-driven synthesis of relevant materials.

Free generative AI systems like ChatGPT (https://openai.com/chatgpt/) and Gemini (https://gemini.google.com/) are well-known for being good sources of inspiration, but their reliability, especially for scholarship where precision is essential, has been repeatedly questioned. For example, the phenomenon of hallucinations extends to sources, with ChatGPT (GPT-3.5) generating either incorrect DOIs or providing the wrong DOI for a source in 61.8% of natural sciences references and 89.4% of humanities ones (Mugaanyi et al., 2024). Hallucinations in AI systems can be divided into one of two types, '*intrinsic hallucinations*' and '*extrinsic hallucinations*' (Mittal et al., 2024). In the first type, intrinsic hallucinations, the output that is generated contradicts the source. In the second type, extrinsic hallucinations, the output is not able to be verified against the source, and is unverifiable.

As mentioned, a new suite of AI-powered tools has emerged, designed to support research. These tools have been organized into systems that help researchers find sources, make sense of primary source texts, and correct their own writing (e.g., Pinzolits, 2023). In addition, literature-based discovery (LBD) is a domain showing much promise. In this model, systems use AI to analyse the literature, specifically seeking lacuna and new approaches. LBD is generating interest in areas such as medicine (e.g., Cheerkoot-Jalim & Khedo, 2021).

Using AI-powered systems to create literature reviews for articles is understudied. Only two studies begin to address the problem. Choe and colleagues (2024) designed a study to assess novice researcher needs in creating reviews of the literature and then created LLM-based scaffolding for the work. Zhao and colleagues (2024) generated literature reviews using three AI-powered document summary systems and one literature review system; they applied Pattern Analysis and Machine Intelligence (PAMI) approaches, comparing AI-generated reviews to human-authored ones using new bibliometric indicators. The results identified limitations in the AI-powered outputs. To our knowledge, no systematic assessment of freely available and specialized AI-powered literature review tools or of their functionalities and their usefulness to academics has been carried out.

Method

Focusing on free and low-cost AI-powered systems that are designed and advertised specifically to support scholars in the creation of literature reviews and that are available to individual researchers (i.e., not requiring enterprise-level commitment by an institution such as with Laser (https://www.laser.ai/)), the research team identified 4 systems for analysis:

- Jenni.ai (https://app.jenni.ai/)
- Scite (https://scite.ai/)
- OpenRead (https://www.openread.academy/)
- Wisio (https://wisio.app/)

Each system was expertly learned by one team member who watched tutorials and practiced extensively, piloting each system. Based on requirements for literature reviews in LIS, a social science (e.g., Machi & McEvoy, 2016; Slater, 2018), and the affordances of AI-powered systems, deductive coding was used for this project.

Paid versions of each system were used to explore maximum benefits, and only the system's sources were used (no sources were uploaded). The goal was a usable 500 to 1000-word literature review on e-reading research in English and Spanish that 1) focused on both languages, 2) cited current and peer-reviewed sources, and 3) summarized methodologies; the research team's system's expert led a live demonstration for the research team beginning with the prompt: '*write a one-page literature review (500-1000 words) about e-reading in Spanish and English.*'

Team leads then modified based on system affordances and results. During the demonstration, which was recorded, members of the research team collaboratively coded each system using an

inductive system based on the requirement of literature reviews in the social sciences, discussing each code until agreement was reached.

Results

Each AI-powered system designed for authoring reviews of the literature was investigated for its ability to produce text for a complex question. Although these systems functioned very differently, the results of their use and observations regarding their feedback, functionality, and responses are presented below.

Table 1 provides an overview of elements relating to authoring a literature review. Specifically, this project sought to produce more than 500 words on the topic, using recent peer-reviewed sources in both English and Spanish, and having the methodologies summarized in the analysis. Scite was the most performant of the group, creating a robust review of the literature using recent sources where methodology was described. None of the systems, however, allowed users to cite only peer-reviewed sources. Jenni.ai and Wisio were the least performant, with Jenni.ai unable to address the majority of the criteria, and only able to isolate methodology in a limited way. Wisio was also unable to address the majority of the criteria, although it was able to cite more recent sources when prompted, but not specifically by date or date range.

System name	>500-word review produced	Specify sources as peer reviewed	Specify sources by language	Specify sources by date	Isolate methodology
Jenni.ai	no	no	no	no	somewhat
Scite	yes	no	no	yes	yes
OpenRead	yes	no	yes	no	somewhat
Wisio	no	no	no	somewhat	no

Table 1. Overview of free and low-cost AI-powered systems for literature reviews.

Table 2 provides an overview of the sources cited and the formatting of the references produced in the test review. As the sources are an essential aspect of the literature review, this table focuses on foundational aspects of how they are provided as well as how the references are given and their accuracy. Again, Scite was the superior product, successfully addressing all the criteria listed. Each product linked to sources in some way or another and none produced hallucinations; Scite provided APA-style formatting and references, and the output text was downloadable as a file that included the formatted references. Jenni.ai likewise met these three criteria, but did not always provide references for sources mentioned in the text. OpenRead and Wisio provided links to sources, but did not provide a standard reference list and did not provide an easy way to download or save outputs. Secondary sources were at times provided in Jenni.ai, OpenRead, and Wisio in lieu of primary sources. The number of references provided ranged from 3 to 10.

System name	Link to sources	Primary sources only	All sources are referenced	Num. references provided	Hallucinations in references	Downloadable output with references	Able to output using APA
Jenni.ai	some	no	no	3	no	yes	yes
Scite	yes	yes	yes	10	no	yes	yes
OpenRead	yes	no	yes	3	no	no	no
Wisio	yes	no	no	10	no	no	no

Table 2. Sources, references, and output produced.

Additionally, scholars could upload sources themselves for inclusion to all four systems, although this project only used sources provided by the systems.

Observations about using the systems

The same prompt was used to begin each interaction with a given. However, the way that each system permitted interaction and prompt refinements differed. Jenni.ai and OpenRead were primarily prompt-based, functioning much like generative AI systems such as ChatGPT and Gemini. After the initial prompt, Jenni.ai allowed sentence-by-sentence approval of content as the primary way to add to the generated content, or for the selection of text around which additional content could be developed. Scite was the most like a traditional database, distilling prompts down to queries and then displaying the queries. Giving OpenRead an identity as a helpful assistant who speaks both languages produced a more scholarly review of literature, though the team observed that the content was not topically aligned with the prompt. Wisio seemed designed to assist in the writing process rather than replace it, with tools to check grammar and to align user input content with scientific writing styles, which Wisio calls '*scientifying*' the writing.

Ability to refine the outputs

Each system allowed for different refinements to the text, via changes to the prompts (e.g., Scite), isolation of specific text to make edits (e.g., Jenni.ai). Features relating to language were difficult to isolate, in all systems producing either no or unpredictable results. For example, Jenni.ai and Wisio were unable to produce a critical mass of references in English, and efforts to request references in Spanish were likewise unsuccessful. Scite was able to change the output language to Spanish, which was unexpected since the prompt was provided in English. In OpenRead, clicking the follow-up option and using the same prompt produced a four-paragraph Spanish-language summary of e-reading research with several APA-style in-text citations. Further prompting was needed before the summary was translated to English and links to references were provided; however, those references were not consistent with the in-text citations.

Scite was the only system that allowed for a date range to be provided for references; however, Jenni.ai, OpenRead, and Scite all provided additional metadata and links to sources. Wisio seemed to use only PubMed and CrossRef to provide citations, but the AI Advisor feature did suggest other databases to search and keywords to use. The overall quality of references varied greatly. For example, OpenRead and Wisio were unable to format references in APA, and OpenRead did not allow for the text of the review to be downloaded with the references appended; instead, each reference needed to be downloaded individually.

Discussion

The results of this investigation provide insight into AI-powered systems for the creation of literature reviews. Based on the performance and functionalities as well as the usefulness of the outputs, we find that the systems performed unevenly.

Overall, based on the requirements of reviews of the literature in the social sciences (Machi & McEvoy, 2016; Slater, 2018), we find that Scite performed admirably by producing text that was both scholarly and well-referenced in alignment with expectations for reviews of the literature (Slater, date). Overall, the Scite output was not that far removed from the kind of output one might expect to find in a published article. Jenni.ai, OpenRead, and Wisio produced sub-par results, even after extensive massaging of the prompts and work within the system. Jenni.ai was unable to produce more than 500 words or include more than 3 references, though it mentioned additional sources in the body of the text. OpenRead produced a paragraph of text with several cited sources, but on review, included journal submission instructions in the content of the text. OpenRead also produced extensive references and scrolling produced more; however, the first 3 references were used in the text and the others were suggestions for related content. Wisio produced partial paragraphs, often ending with unfinished sentences. References provided were not specific to the

prompt, though Wisio's AI Advisor feature did describe and cite one study that met requirements for timeliness and that seemed to be on-topic.

All of the systems allowed for sources to be loaded and used. Uploading articles under copyright for use by these systems is possibly questionable in terms of intellectual property rights. Accordingly, this project sought to understand the affordances of the AI-powered systems without such inputs; however, for users who wanted to focus on specific sources, the ability to upload those sources might be a useful approach.

Limitations and future research

This exploratory study does not seek to answer questions about e-reading in English and Spanish empirically, nor to weigh in on the ethics of the design, functionality, or use of such systems in academic writing, but rather to compare the functionalities of free and low-cost AI-powered systems for creating literature reviews. The research questions investigated the nature of using the systems and the mechanics of the output, rather than the more subjective veracity of the contents. Further research will investigate the accuracy of the text produced, the relevance of the sources cited, and the ability of the AI to identify the most important and pressing concerns relating to the topic. As AI advances, further research should also continue to monitor the strengths and weaknesses of these and other systems.

Conclusion

AI-powered systems show incredible promise, with one unstudied area being the creation of scholarly reviews of the literature. Although the kinds of research questions that might be examined in social sciences like LIS will necessarily be complex, until now little has been done to evaluate these systems for their potential use in scholarship and teaching. This paper is a first attempt to systematically analyze the functionalities of four AI-powered systems for reviews of the literature that are available for use by individuals.

About the authors

Heather Moulaison-Sandy is Associate Professor in the iSchool, University of Missouri, USA. She received her Ph.D. from Rutgers University, and her research interests focus on the organization of information, scholarly communication, and technology. She can be contacted at moulaisonhe@missouri.edu

Wilson Castaño-Muñoz is Associate Professor at the School of Library and Information Science, University Antioquia, Colombia. He is a PhD student at the University of Missouri. His research interests are in digital reading and information conveyed through aural technologies. He can be contacted at wilson.castano@mail.missouri.edu

Laura Ridenour is Assistant Professor in the iSchool, University of Missouri, USA. She received her Ph.D. from the University of Wisconsin-Milwaukee, and her research interests include data science, artificial intelligence, information organization, and open data governance and access. They can be contacted at lerbhn@missouri.edu

Denice Adkins is Professor in the School of Information Science & Learning Technologies, University of Missouri. She received their Ph.D. from the University of Arizona and her research interests are in services to library users. She can be contacted at adkinsde@missouri.edu

References

Antu, S. A., Chen, H., & Richards, C. K. (2023). Using LLM (Large Language Model) to improve efficiency in literature review for undergraduate research. LLM@ AIED, pp. 8-16.

Bolter, J. D., & Grusin, R. A. (1996). Remediation. Configurations, 4(3), 311-358.

Canclini, N., Gerber Bicecci, V., & López Ojeda, A. (2015). Hacia una antropología de los lectores. D - Ediciones Culturales Paidós. México.

Cavallo G., & Chartier R. (Eds.). (1999). A history of reading in the west. Amherst: University of Massachusetts Press.

Cheerkoot-Jalim, S., & Khedo, K. K. (2021). Literature-based discovery approaches for evidence-based healthcare: A systematic review. Health and Technology, 11(6), 1205–1217. https://doi.org/10.1007/s12553-021-00605-y

Choe, K., Park, S., Jung, S., Kim, H., Yang, J. W., Hong, H., & Seo, J. (2024). Supporting novice researchers to write literature review using language models. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–9. https://doi.org/10.1145/3613905.3650787

Consensus AI. (2022, September 14). Consensus and Semantic Scholar Partner for Approachable Information. https://consensus.app/home/blog/science-answer-app-consensus-and-article-aggregator-and-discovery-tool-semantic-scholar-announce-their-partnership/

Cordón, J. A. (2016). La lectura en el entorno digital: nuevas materialidades y prácticas discursivas. Revista chilena de literatura, (94), 15-38. https://www.scielo.cl/pdf/rchilite/n94/art02.pdf

ijmarkandel. (2023). AI tools for literature review. r/PhD. Reddit. https://www.reddit.com/r/PhD/comments/17nqf59/ai_tools_for_literature_review/?rdt=52038

Joos, L., Keim, D. A., & Fischer, M. T. (2024). Cutting through the clutter: The potential of LLMs for efficient filtration in systematic literature reviews. *arXiv Preprint arXiv:2407.10652*. https://arxiv.org/pdf/2407.10652

Machi, L. A. & McEvoy, B. T. (2016). The literature review: Six steps to success. 3rd ed. Thousand Oaks, California: Corwin.

Mittal, A., Murthy, R., Kumar, V., & Bhat, R. (2024, January). Towards understanding and mitigating the hallucinations in NLP and Speech. In: *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pp. 489-492.

Morales, O., & Espinoza, N. (2003). Lectura y escritura: coexistencia entre lo impreso y lo electrónico. Educere, 7(22), 213-222.

Mugaanyi, J., Cai, L., Cheng, S., Lu, C., & Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. Journal of Medical Internet Research, 26, e52935. https://doi.org/10.2196/52935

Pinzolits, R. F. J. (2023). AI in academia: An overview of selected tools and their areas of application. MAP Education and Humanities, 4(1), 37–50. https://doi.org/10.53880/2744-2373.2023.4.37

Shopovski, J. (2024). Generative artificial intelligence, AI for scientific writing: A literature review. Preprints, 2024060011. https://doi.org/10.20944/preprints202406.0011.v1

Slater, T. F. (2018). Improving your argument by identifying a literature gap. Journal of Astronomy and Earth Sciences Education, 5(1), i–ii.

Zhao, P., Zhang, X., Cheng, M. M., Yang, J., & Li, X. (2024). A literature review of literature reviews in pattern analysis and machine intelligence. arXiv preprint arXiv:2402.12928.

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Get the root of research misconduct: a systematic review of its cultural factors

Wei Feng, Lihong Zhou and Junmin Xiao

DOI: https://doi.org/10.47989/ir30iConf46903

Abstract

Introduction. Culture significantly shapes scientists' perceptions of research misconduct (RM) since RM can differ from culture to culture. Yet the absence of a comprehensive understanding of cultural factors affecting RM often results in the oversight of cultural influences when implementing RM-preventing measures.

Method. This study adopted a systematic review approach to identify specific cultural factors of RM and conceptualize their influencing relationships with RM.

Analysis. Guided by cultural dimension theory, a total of 55 articles were analysed through thematic analysis.

Results. Five primary cultural factors emerged: strong authority hierarchy, collectivist culture, self-interested culture, gender roles, and low uncertainty tolerance. These cultural factors, serving as common origins of other typologies of RM factors, were found to indirectly influence RM by interacting with other factors.

Conclusions. This review contributes to the theoretical expansion of knowledge on the cultural factors of RM and a conceptual model to understand their relationships with RM. The findings can provide useful implications for developing effective and targeted strategies to mitigate the negative effect of cultural factors on RM.

Introduction

Nowadays, the evolving and uncertain academic landscape, particularly with the emergence of artificial intelligence (AI) technologies like ChatGPT, requires library and information science (LIS) professionals to environmentally scan factors affecting research misconduct (RM) and extend their roles in research integrity (RI) (Auer & Krupar, 2001; Benjes-Small et al., 2008; Caravello, 2008; Herr, 2019; Seadle, 2023). However, AI does not inherently induce individuals to engage in RM; rather, it merely facilitates the process. It is people's behaviours such as ethical decision-making, cognitive processes, and perceptions of societal values that ultimately determine whether they commit RM or not (Li & Cornelis, 2021; Olesen et al., 2017). Notably, people's behaviours in RM are oriented by various factors but fundamentally shaped by culture (Durham, 1976; Eckstein, 1997; Lonner & Adamopoulos, 1997; Valkenburg et al., 2021).

Culture is a highly intricate concept that remains the subject of considerable debate (G. J. Hofstede, 2015). Briefly, culture is characterized as '*the customary beliefs, social forms, and material traits of a racial, religious, or social group*' (Merriam-Webster, n.d.). Culture plays a significant role in determining what is considered acceptable or unacceptable behaviour within a society. Relating more specifically to research settings, the judgement and acceptance of RM can differ from culture to culture (Farahat, 2022; JAL, n.d.). Thus, scientists might commit RM intentionally driven by unhealthy cultural traits or unintentionally due to unawareness of RI principles in other cultural backgrounds.

To address cultural influences on RM, certain measures on education and training are advocated to improve the ethical awareness of scientists (Rodrigues et al., 2023). Beyond education, cultural reforms are implemented to tentatively drag scientists back to the pure intention of embarking on scientific careers (Casadevall & Fang, 2012). Additionally, some governance models of cultural reforms are proposed to handle RM (Schiermeier, 2010; Schocker et al., 2021). However, these measures merely play a marginal role in addressing the cultural challenges of RM due to the lack of deep understanding of the specific cultural factors.

There is now an emerging body of work on exploring the factors of RM, primarily encompassing three typologies: individual factors, organizational factors, and professional factors (Feng et al., 2024; Fierz et al., 2014; Haven & van Woudenberg, 2021; Horbach et al., 2019; Wang & Zhu, 2023). While existing studies have explored the impact of cultural norms, values, and practices on RM (Buljan et al., 2018; Davis, 2003; Rodrigues et al., 2023), culture has been merely studied as an explanatory lens for RM rather than as a typology of direct or indirect factors. There is still a lack of comprehensive understanding and in-depth analysis regarding specific cultural factors, their interaction with other typologies of RM factors, and the mechanism of how they shape the attitudes of scientists towards RM.

To bridge these gaps in RM studies, the present study aims to delve deeper into the specific cultural factors of RM beyond superficial observations and explanations. Theoretically, the analysis of the present study contributes to a more nuanced understanding of cultural factors affecting RM and their relationships with RM. Practically, the cultural-factor map of RM developed by the present study can be utilized by interest groups to navigate the complex cultural landscape and thus to work towards implementing effective and targeted strategies. Hence, the following research questions are of significant concern in the present study:

- What are the cultural factors that influence RM?
- What are the relationships between RM and its cultural factors?

Literature review and theoretical basis

Research misconduct

Based on an overview of RM definitions issued by different research institutions across various research disciplines (Gilbert & Denison, 2003), early definitions tend to be inclusive and generic, not specifying what is RM and what is not. For instance, in the 1990s, the US National Science Foundation simply defined RM as serious deviations from generally accepted practices in a particular scientific discipline (Bornmann, 2013; Buzzelli, 1993). This kind of definition leaves more room and freedom to work against all potential forms of aberrant conduct (Bornmann, 2013).

The most widely used definition, coined by the Office of Science and Technology Policy (OSTP) of the US White House (The Office of Science and Technology Policy, 2000), refers to '*fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results*'. This definition specifically points out what is not acceptable, and what is an offence against the rules of good scientific practice. In this definition, fabrication is making up data, processes, and results and claiming them as real. Falsification means the manipulation of research data, materials, equipment, or processes to yield a better research outcome. Plagiarism refers to untruthful reports of research processes and results or using other people's work or ideas without acknowledging the source.

In recent years, the international academic community has observed a surge in instances of RM (Martinson et al., 2005; Wang & Zhu, 2023). The persistent occurrence of RM has significantly undermined the credibility and value of scientific research, eroded public trust in the scientific enterprise, and compromised the motivation and morale of scientists to engage in innovative endeavours (Martinson et al., 2005; Nelson & Lubchenco, 2022; Seadle, 2017; Xie et al., 2021). To address the issues of RM, several studies have paid attention to preventing, detecting, correcting, and reporting RM (Bonito et al., 2012; Buljan et al., 2018; Gillespie, 1991; Gunsalus, 2019; Kumar, 2010; Wiedermann, 2018), as well as cultivating and repairing RI (Mejlgaard et al., 2020; Tang, 2019; Titus et al., 2008). Specifically, numerous measures have been implemented, such as integrity training and education, targeting fake peer reviews, strengthening management of science foundations, and providing guidance and policy (Cyranoski, 2017; Y. Gong, 2005; Mejlgaard et al., 2020; Michel, 2010; Yi et al., 2019).

Factors of research misconduct

The measures of preventing RM are often developed without a deep understanding of the factors of RM, which are characterized by '*secrecy and complexity*' in research (DuBois et al., 2013, p. 321). A comprehensive understanding of the underlying factors of RM can facilitate the prediction of potential RM, the design of targeted interventions, and the formulation of policy decisions and educational initiatives, among other benefits (Fanelli et al., 2015; Farahat, 2022; Horbach et al., 2020; Horbach & Halffman, 2019; Mardani et al., 2020).

As reported in the existing studies, RM factors can range from individual personality traits, and organizational characteristics to professional causes (Committee on Responsible Science et al., 2017; Fanelli et al., 2022; Feng et al., 2024; Franzen et al., 2007; Horbach et al., 2019; Sovacool, 2008). Some researchers map these three typologies of RM factors to micro, meso, and macro levels respectively, indicating that the factor of RM is a multilevel problem (Faria, 2015; Fierz et al., 2014; Mardani et al., 2020; Schocker et al., 2021). Individual narrative of RM factors suggests the existence of '*scoundrels among scientists*' (Sovacool, 2008, p. 276), informing that individual personality traits such as narcissism, overconfidence, and a lack of ethical values can be major factor of RM (Cairns et al., 2021; Krstić, 2015; Tijdink et al., 2016). According to Cairns et al. (2021), Mardani et al. (2020), and Tourish and Craig (2020), organizational factors such as institutional pressures, leadership style, and organizational culture can lead to RM. Additionally, professional factors stem from '*professional competitions and pressures among scientists*' (Feng et al., 2024, p. 7). They exist across

organizations and throw threats to the whole academic system, including publication bias, the current research funding system, the pressure to publish (Buljan et al., 2018; Golden et al., 2023; Gopalakrishna et al., 2022; Horbach & Halffman, 2019; Rodrigues et al., 2023), etc. The specific RM factors of these three main typologies, according to Feng et al. (2024), are shown in Table 1.

Themes	Sub-themes (Specific Factors of RM)
Individual Factors	Personal characters
	Lack of ethical awareness
	Lack of experience
	Unstable mental state
	Excessive personal and family pressures
Organizational Factors	Insufficient RI education and training
	Lack of explicit rules and regulations
	Lack of supervision and mentoring
	Lack of punishment and insufficient censure
	Lack of report and transparency on RM
	Misdirected reward system
Professional Factors	'Publish or perish' pressure
	Pressure from grant application
	Pressure from promotion and tenure
	Overly intense peer competition

Table 1. Individual, organizational, and professional factors of RM

Besides the above three typologies of RM factors, Davis (2003) developed a conceptual model of RM factors supplementing cultural factors and situational factors. The five typologies of RM factors in Davis's model have an overlapping and parallel relationship with each other as shown in Figure 1. Davis's (2003) model suggested that diverse typologies of RM factors can take effect together on RM. Situational factors refer to stressful and extrinsic circumstances that can encompass situational challenges to scientists and easily induce them to commit RM (Committee on Responsible Science et al., 2017). Since situational factors can be essentially traced back to professional competitions and pressures among scientists (Davis et al., 2007; Faria, 2015), it can be recognized as a combination of organizational factors (i.e. organizational situation) and professional factors (i.e. professional situation) to some extent.

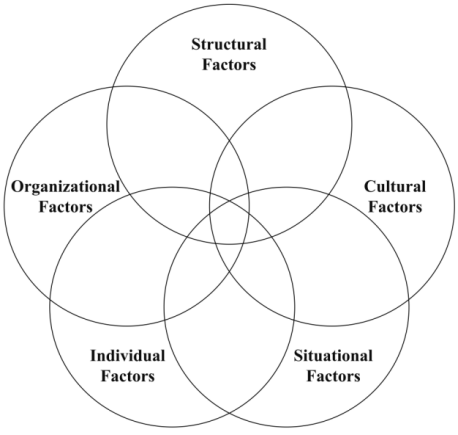


Figure 1. The conceptual model of RM factors proposed by Davis (2003)

Davis (2003) brought culture into the research landscape of RM factors, suggesting that scientists may engage in RM in some ways that they may find appropriate but that are not appropriate in other cultures. Factors such as attitudes towards competition, success, and hierarchy (Aubert

Bonn & Pinxten, 2021; Buljan et al., 2018), as well as the emphasis on individual achievement versus collective well-being (Mirshekary & Lawrence, 2009; Parnther, 2022), can contribute to the occurrence of RM and have roots in culture. The influence of culture on RM has been studied with diverse research methods, such as the comparative approach to understand its universality and specificity (Stephens et al., 2010; Williams et al., 2014), literature analysis to construct a comprehensive landscape of cultural contexts (Rodrigues et al., 2023), and theoretical discussions to expand RI from the cultural-practice perspective (Valkenburg et al., 2021). Culture may influence people's behaviours in RM by predisposing them to certain actions, shaping their values, influencing their knowledge acquisition, and structuring distributions of responsibility and accountability (Valkenburg et al., 2021). Despite significant studies exploring the cultural influence on RM, the whole research landscape of cultural factors affecting RM is still fragmentary and obscure. Furthermore, the influencing mechanism between cultural factors and RM behaviours remains underexplored.

Some established cultural theories have been used to explain the cultural factors of RM. For instance, Sellin's (1938) theory of culture conflict has been employed to explicate the factors for the divergent ethical standards among researchers from different nations when facing similar RM issues (Davis et al., 2007; Jesilow et al., 1992; Parnther, 2022). As well, Merton's (1938) theory of social structure and anomie posits that scientists are inclined to engage in RM when societies excessively prioritize socially desirable outcomes (e.g., economic growth) over reasonable and lawful means. For instance, the '*winner-take-all system*' in academia prioritizes breakthroughs in scientific discovery, contributing to an atmosphere in which some scientists attempt to commit RM (Casadevall & Fang, 2012, p. 892; Stern et al., 2014). Nevertheless, these theories only offer one-sided interpretations for the cultural factors affecting RM and does not clearly clarify the influencing mechanism of cultural factors on RM.

Cultural dimension theory as a theoretical lens

To conceptualize these cultural factors of RM and comprehend the encountered cultural challenges, G. Hofstede's cultural dimension theory is adopted as a theoretical lens. This choice is motivated by the cultural dimension theory's (a) stability over time (G. J. Hofstede, 2015), (b) prevalence across various national cultures globally albeit to varying degrees (G. J. Hofstede, 2015), and (c) application and effectiveness in numerous RM studies to explore potential factors (e.g., Mirshekary & Lawrence, 2009; Olesen et al., 2017; Parnther, 2022; Williams et al., 2014). The initial framework of cultural dimension theory only contained four dimensions (G. Hofstede, 1984). In response to critiques of Western bias, G. Hofstede and Bond (1984) introduced a fifth dimension to specifically address Eastern cultural traits, making the cultural framework highly relevant to the RM issues globally addressed in this paper (Zhou & Nunes, 2013). Until 2010, G. Hofstede had added a sixth dimension to the cultural dimension theory with the cooperation of Michael Minkov (G. Hofstede, 2011; G. Hofstede et al., 2010).

The latest cultural dimension theory consists of the following six dimensions:

- *Power distance* reflects how countries '*handle the fact that people are unequal*' (G. Hofstede et al., 2010, p. 55). It refers to the degree to which less powerful individuals accept and anticipate unequal distribution of power (G. Hofstede, 1994). Societies with high levels of power distance tend to have rigid hierarchies, a strong respect for authority, and a reluctance to challenge those in positions of power.
- *Individualism versus Collectivism* is the degree of integration into groups. Individualism has loose ties between individuals and an expectation of self-care and immediate family. Collectivism has strong in-group ties from birth, often extended families, offering protection for loyalty (G. Hofstede, 1994; G. Hofstede et al., 2010, p. 92).

- *Masculinity versus Femininity*: in masculine societies, where '*emotional gender roles are clearly distinct*' (G. Hofstede et al., 2010, p. 140), men's values vary from being assertive and competitive to modest and caring, with a contrast between men's and women's values. Feminine societies, where '*emotional gender roles overlap*' (G. Hofstede et al., 2010, p. 140), have comparable modest and caring values for both genders.
- *Uncertainty Avoidance* describes a society's tolerance for uncertainty. It reflects the extent to which a culture programs its members to feel comfortable or uncomfortable with ambiguity, uncertainty, and risk (G. Hofstede et al., 2010, p. 140).
- *Long- versus Short-Term Orientation* deals with *Virtue regardless of Truth* (G. Hofstede, 1994, p. 5). Long-term orientation values thrift and perseverance while short-term orientation values tradition, social obligations, and '*face*' protection (G. Hofstede et al., 2010, p. 237).
- *Indulgence versus Restraint* is *'related to the gratification versus control of basic human desires related to enjoying life*' (G. Hofstede, 2011, p. 8). The indulgence represents freedom and enjoyment, associated with high happiness, while the opposite reflects restriction and social norms, leading to a feeling of inappropriateness (G. Hofstede et al., 2010, p. 281).

It is worth noting that certain existing studies (Coseglia et al., 2023; Faria, 2015; Haven & van Woudenberg, 2021) have included working or organizational culture in the category of cultural factors affecting RM. However, working, or organizational culture primarily and directly contributes to external factors (e.g., organizational factors, professional factors) rather than internal motivations of people's behaviours. On the other hand, working or organizational culture represents a situational '*group identity*' that changes when the working environment changes rather than a typology of '*culture*' (G. J. Hofstede, 2015, p. 224). Hence, this study does not consider working or organizational culture as the primary cultural factors. Nevertheless, if certain kinds of working or organizational culture are rooted in the cultural dimensions outlined in G. Hofstede's theory, it is still worthy of discussion in the present study.

Research methodology

The preliminary review developed theoretical and contextual indications to guide the remaining research stages. It not only confirmed the research questions and research objects but also suggested that the influence of culture on RM should be present, although it is constrained (Davis, 2003; Feng et al., 2024; Mirshekary & Lawrence, 2009; Olesen et al., 2017). This research aims to construct a comprehensive landscape of factors affecting RM and conceptualize their influencing relationships with RM. RM is a complex field and investigations into its cultural factors employ diverse research methods. Mixed studies should be included in the review process to harness the advantages of both quantitative and qualitative evidence (Hong & Pluye, 2018; Petticrew et al., 2013). Thus, a systematic review approach was adopted to "systematically search for, appraise and synthesis research evidence" as suggested by Grant and Booth (2009, p. 102).

Whilst the preliminary review was designed to be general and wide, the research team intended to follow the systematic review procedures and perform a rigorous analysis focusing on the cultural factors of RM during the main review process. Two international databases, Web of Science Core Collection and Scopus, were employed for literature retrieval. These databases jointly cover the core collection of peer-reviewed research articles published across different disciplines.

Nine studies from the preliminary review (Section 2.2) were first used as references for the retrieval strategy establishment and included in the main review process. The retrieval strategy consisted of two parts for pairing searching: one was about RM, including its synonyms such as academic misconduct and research misbehaviour, as well as several narrower keywords like fabrication, falsification, plagiarism, and academic dishonesty. The other was about "cultural factors", including its synonyms like cultural determinants, cultural contributors, cultural causes, and so on.

The retrieval strategy was designed to be inclusive of as many relevant articles as possible. To avoid unnecessary dilemmas in the RM definition, the retrieval strategy rigorously adopted the OSTP definition. Given the embedded nature and correspondence of culture in dishonest behaviours proposed by Casadevall and Fang (2012), academic dishonesty behaviours were included when searching relevant articles. The search process was completed between November and December 2023, not restricting the publishing year of articles. The idea is to create a sufficient pool of articles, from which a manual search can be performed.

The database retrieval returned a total of 440 articles after removing duplicated records based on their titles, authors, and publishing year. The research team formed a series of inclusion criteria to guide the identification and screening process. For inclusion, a study must meet the following criteria: (a) it was published in English; (b) it is one of the journal articles or conference proceeding articles; (c) it has been peer-reviewed before publication; (d) its full text is available for download; (e) it must report cultural factors of RM; (f) it must conform the OSTP definition of RM. Adhering to these inclusion criteria, a total of 37 articles were included during the database-retrieval stage.

Thereafter, citation tracking was conducted manually on the included 46 articles (9 from preliminary review and 37 from database retrieval) through Google Scholar, which allowed the identification of relevant articles not obtained from the previous stages. Using this approach, another nine articles were included for analysis after identification and screening with the same inclusion criteria. Finally, a total of 55 relevant articles were included for systematic review analysis. The whole process of selecting articles was demonstrated in the PRISMA diagram (Figure 2) as suggested by Page et al. (2021).

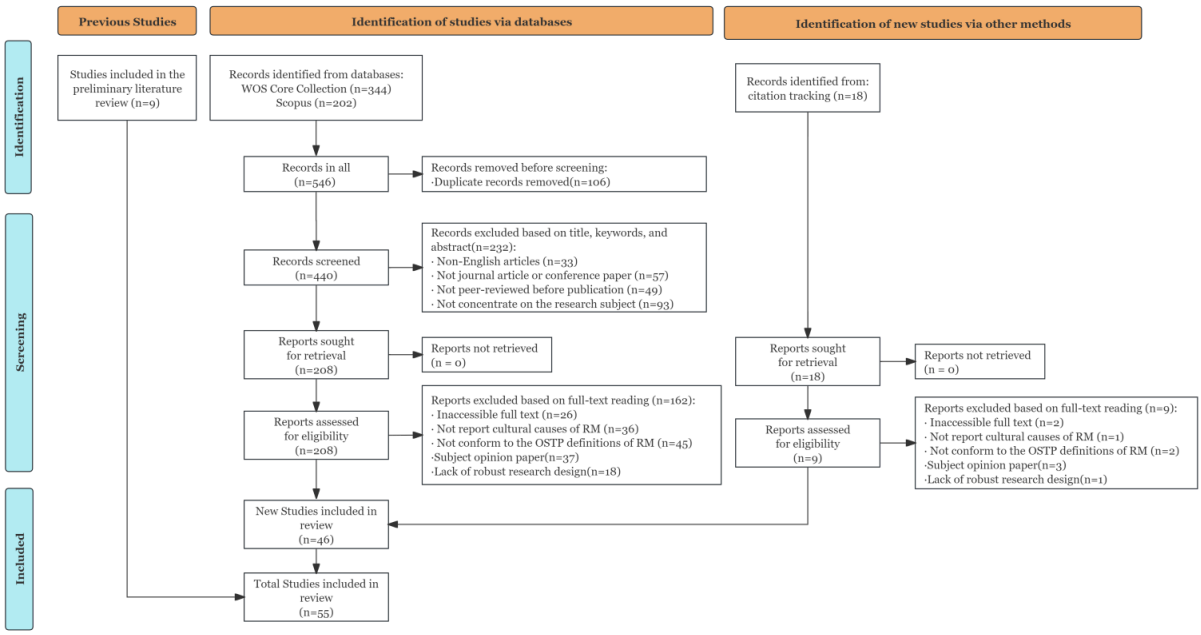


Figure 2. PRISMA diagram on article screening processes

The method of thematic analysis was adopted to analyse selected articles, with the cultural dimension theory discussed in Section 2.3 used to guide, orient, and frame the practice of coding. Two authors coded the nine articles from the preliminary review firstly with discussions to formulate the coding agreement and reserved the nine articles from citation tracking in advance for theoretical saturation tests. The full text of all selected articles was imported into MAXQDA Analytics 2020 for inductive coding. In the coding process, the pre-set list of codes was continually examined, confirmed, and updated while new codes were anticipated to emerge from the literature (Saunders et al., 2009). Moreover, three coding

techniques were used, namely open coding, axial coding, and selective coding (Strauss & Corbin, 1998).

Research findings

The literature analysis pointed to five main cultural factors of RM: strong authority hierarchy, collectivist culture, self-interested culture, gender roles, and low uncertainty tolerance. The themes and codes that emerged from the thematic analysis process are shown in Table 2.

Themes (Cultural Factors)	Codes (Manifestations of Cultural Factors)
Strong authority hierarchy	·Pressure to deference ·Fear of challenging authority ·Limited transparency ·Lack of accountability
Collectivist culture	·Loyalty to peers and affiliated institutions ·Safeguarding the collective image ·Persistence to norms of home countries
Self-Interested culture	·Diluted responsibility among collaborators ·Pursuit of personal interest
Gender roles	·More RM cases in male scientists than in females. ·Academic society as a masculine society
Low uncertainty tolerance	·Reluctance to innovation ·Avoidance of work insecurity ·Education practices emphasizing memorization

Table 2. Themes and codes of thematic analysis

Strong authority hierarchy

The risk of RM is higher in countries that are more hierarchical and less regulated (Fanelli et al., 2015, 2017). In terms of geographic distribution, RM cases are more often reported in some Asian countries like China, Indonesia, and India (Aubert Bonn & Pinxten, 2019; Putra et al., 2023; Rodrigues et al., 2023), which demonstrate high authority hierarchy cultural traits (G. Hofstede et al., 2010, p. 60). The literature analysis revealed some specific manifestations of strong authority hierarchy leading to RM, including pressure to deference, fear of challenging authority, limited transparency, lack of accountability, etc.

Strong authority hierarchy is closely related to a culture of deference to those in positions of authority (G. Hofstede et al., 2010, p. 67), such as principal investigators and senior researchers in academia (Yu et al., 2021). Subordinates working in strongly hierarchical structures are more likely to trust and obey their leaders, as well as recognize their leaders' views and behaviours (Z. Gong et al., 2019; Graham et al., 2018; Peltokorpi, 2019). This has led some students to blindly believe that the information provided by their supervisors is entirely accurate, then directly utilizing this information without critical verification and consequently engaging in RM. The notorious case of Diederik Stapel serves as a classic example of this phenomenon (Retraction Watch, 2016). Furthermore, reports of RM by junior scholars are less likely to be spoken up, let alone be effective in the cultural environment of strong authority hierarchy, as they may fear negative consequences for challenging those with stronger authority (Buljan et al., 2018; Horbach et al., 2020). Limited transparency and openness on RM will therefore be intensified, making it easier for RM to occur without intervention.

More seriously, power-holders in a situation with strong authority hierarchy even have the privilege to, by deliberately misrepresenting information, dictate what is plagiarism, whether plagiarism belongs to the category of RM or not, as well as who should be punished due to RM (Horbach et al., 2019; Putra et al., 2023). In this regard, the lack of accountability and oversight in

the research process will happen since those in positions of authority are less likely to be held accountable for unethical practices (Velayutham & Perera, 2004).

Collectivist culture

The collectivist culture has a significant impact on academic behaviours (Farahat, 2022; Mirshekary & Lawrence, 2009; Parnther, 2022). In collectivist cultures, researchers are connected with some critical social elements like trust, confidence, and cooperation, which are conducive to the emergence of RM behaviours (D. W. Chapman & Lindner, 2016; Tourish & Craig, 2020). Some collectivist cultures prioritize loyalty to peers and affiliated institutions as more important than compliance with RI principles (Bertram Gallant, 2015). This cultural trait has led to an environment where RM is facilitated by institutional protection of wrongdoers and informants' reluctance to divulge any information about RM for the safeguarding of the collective image (Schulz, 2015; Tourish & Craig, 2020). Empirical evidence from Williams et al. (2014) supported this viewpoint by demonstrating that students from collectivist cultures (e.g., the United Arab Emirates and some other nations from the Middle East) were less inclined to take RM as serious violations compared to those from individualist cultures (e.g., the United States).

The diverse cultural landscape within and between countries contributes to varying perceptions of ethically troubling behaviours (Fonseca et al., 2016). Specifically, foreign researchers often face distinct pressures to deviate from established norms in their home society, which do not take some behaviours as RM (Davis, 2003; Davis et al., 2007). This can be attributed to the belief of foreign researchers in the values of their home society out of a collectivist mindset (Parnther, 2022).

However, collectivist culture presents a dual impact on RM. Scientists in collectivist environments exhibit higher levels of self-esteem (Rehman & Waheed, 2014), which is advantageous for RI due to their confidence in research and the desire to protect their reputation and that of their groups (Tijdink et al., 2016). Additionally, the incidence of RM may decrease with an increasing number of co-authors in a collectivist culture due to enhanced self-control over the content and origin of research (Horbach & Halffman, 2019).

Self-interested culture

The self-interested culture has emerged from literature analysis as an important cultural factor to perpetuate RM behaviours. Scientists in the individualist pole are inclined to do some challenging work including RM to satisfy '*a personal sense of accomplishment*' (G. Hofstede et al., 2010, p. 92), especially when the benefits from RM outweigh the loss (Tourish & Craig, 2020).

Completely different from collectivist cultures, in self-interested cultures, the occurrence of RM can rise when the average number of co-authors on a publication increases since responsibility among authors becomes diluted (Bennett & Taylor, 2003; Horbach & Halffman, 2019; Sun, 2013). Self-interested researchers, who often prevail in placing individual interest above collective interest (G. Hofstede et al., 2010, p. 130), may engage in RM for personal gain disregarding the reputation of their collaborators and their collective (Committee on Responsible Science et al., 2017; Gillespie, 1991; G. Hofstede, 1984; Tourish & Craig, 2020). Contrary to the popular opinion and the stereotype that collectivists like Asians are more likely to plagiarize due to their cultural norms, studies conducted by Kasler et al. (2021) and Martin (2012) support the notion that individualists are more prone to plagiarism due to their pursuit of personal interest.

Gender roles

The thematic analysis demonstrated that gender roles could play an impact on RM behaviours (Aubert Bonn & Pinxten, 2019; Fanelli et al., 2022; Rodrigues et al., 2023). Academia is characterized as a classically masculine society (Bagilhole, 2002; Fotaki, 2013), in which the aggressive, competitive, and assertive psychological characteristics of male scientists predispose them to seek status and take risks through RM (Fanelli et al., 2015, 2017). Data from the US Office of Research

Integrity indicated a higher incidence of RM cases involving male scientists compared to their female counterparts (Fang et al., 2013). Other studies (e.g., Ghias et al., 2014; Gopalakrishna et al., 2022; Mirshekary & Lawrence, 2009; Yu et al., 2021) also illustrated a greater tendency for male scientists or students to engage in RM behaviours. Kaatz et al. (2013) attributed this phenomenon to the implicit gender schema, which potentially contributed to a lower threshold for male scientists to engage in RM, influenced RM case investigations, and affected the distribution of funding between male and female scientists.

While female scientists statistically exhibit fewer RM cases than males, this does not negate the potential influence of femininity on RM. Conversely, feminine cultural traits may influence a collaborative and nurturing approach (G. Hofstede et al., 2010, p. 139), which could contribute to authorship issues such as ghost, guest, or gift authorship during the process of information dissemination (Rodrigues et al., 2023; Smith et al., 2020; Yu et al., 2021).

Although gender can play a role in RM, the statistical relationship between gender roles and RM is not established universally across scientific disciplines (Dannhoferová et al., 2022; Fanelli et al., 2015). Males continue to dominate a significant portion of the academic society and power structure today, suggesting that the link between gender roles and RM may be '*modulated by career or status*' (Fanelli et al., 2015, p. 11). The extent to which gender role contributes to RM also varies across countries, indicating the need for further research in this area (Farahat, 2022).

Low uncertainty tolerance

The literature analysis revealed that nations with low uncertainty tolerance have a high acceptance of RM behaviours. Several comparative studies (e.g., Chapman & Lupton, 2004; Salter et al., 2001) between different countries supported the impact of low uncertainty tolerance on RM. Iranian students, for instance, displayed a greater acceptance of serious RM compared to Australian students. This is reflected in Iran's highest dimension ranking of 'uncertainty avoidance' at 59, indicating a lower tolerance for uncertainty, while Australia's score was 51, indicating a higher tolerance for uncertainty (Mirshekary & Lawrence, 2009). Individuals in societies with low uncertainty tolerance tend to take actions to ensure certainty and are easily caught into "high stress and high anxiety" (G. Hofstede et al., 2010, p. 203), which contribute to RM mainly through certain ways emerged from literature analysis, including reluctance to innovation, avoidance to work insecurity, education practices emphasizing memorization, etc.

The low uncertainty tolerance culture tends to resist the innovative process in organizations (G. J. Hofstede, 2000), as it creates an environment where individuals are uncomfortable with ambiguity and prefer established routines and procedures. This resistance to change and uncertainty can lead to a reluctance to challenge established norms and practices, as people may fear the potential negative consequences or failure that could arise from trying something new (G. Hofstede et al., 2010, p. 169). Consequently, researchers working in such cultures may find it challenging to engage in innovative research (Jam et al., 2018; Kaasa & Vadi, 2010), as they are often discouraged from taking risks or deviating from the status quo (Kaasa, 2016). This fear of negative outcomes and the pressure to conform can sometimes push researchers towards engaging in RM practices, such as data manipulation and fabrication (Kaasa & Vadi, 2010). However, the direct link between the fear of reporting negative results and the engagement in RM is not well-established and requires further investigation.

Additionally, job insecurity is a typical performance of low uncertainty tolerance (Jam et al., 2018; Roll et al., 2015). It, in scientific work, is viewed as a trigger for RM (Davis et al., 2007) because individuals with a low tolerance for uncertainty prefer structured and secure lives, and may feel threatened by job insecurity. Such individuals thus are more likely to resort to RM to enhance their academic performance and reduce insecurity (Afsar & Masood, 2018; Lee et al., 2018; Staufenbiel & König, 2010), especially when the academic society is becoming hypercompetitive.

Education practices prioritizing memorization, replication, and rote learning are accepted by some national cultures like China (G. Hofstede, 1986), which reflect a pursuit of stability and norms, consistent with the cultural trait of low uncertainty tolerance. Although the ability to memorize lectures or source material is often awarded (Madray, 2013), these education practices are admitted to be one of the major factors of plagiarism (Hu & Lei, 2012; James et al., 2019; Rodrigues et al., 2023; Shi, 2006). Scientists who were raised in these education practices are often less tolerant of uncertainty and, therefore, more likely to engage in plagiarism (Dannhoferová et al., 2022; Rodrigues et al., 2023; Vučković et al., 2020). On the one hand, individuals from cultures that rely on memorization are more likely to condone RM behaviours like unattributed information sources in papers due to the lack of critical thinking (James et al., 2019; O'Sullivan & Guo, 2011). On the other hand, replication of original information, instead of being regarded as plagiarism, is sometimes accepted by certain cultures with low uncertainty tolerance as respect for original work, while rational synthesis may be classified into the queue of disrespect (Gregson & Kell, 2009).

Discussion

The thematic analysis of this systematic review identifies five main cultural factors: strong authority hierarchy, collectivist culture, self-interested culture, gender roles, and low uncertainty tolerance. These five cultural factors align with high power distance, collectivism versus individualism, masculinity versus femininity, and high uncertainty avoidance respectively in G. Hofstele's theory of cultural dimensions. It was observed that certain cultural dimensions (i.e., *High Power Distance*, *High Uncertainty Avoidance*) were unidirectional, with one aspect having an impact. Some cultural dimensions (i.e., *Collectivism versus Individualism*, *Masculinity versus Femininity*) exhibited bidirectional associations with both aspects potentially affecting RM. Thus, G. Hofstede's theory of cultural dimensions can provide plausible explanations for the influence of cultural factors on RM (Zhou & Nunes, 2013). For instance, traditional masculinity norms emphasize competition, dominance, and the pursuit of recognition and prestige (Hofstede et al., 2010, pp. 137-138), fostering an environment where scientists are more inclined to achieve impressive results through RM.

While the cultural dimension of *Long- versus Short-Term Orientation* did not receive support from the selected articles in this systematic review, this does not negate the potential association of cultural characteristics represented by this dimension with RM. For instance, values such as "freedom, rights, achievement, and thinking for oneself" (G. Hofstede & Minkov, 2010, p. 497) are excessively emphasized in short-term-oriented cultures. Within this cultural context, the state or research institutions may employ short-term evaluation methods that tie financial rewards to publications, thereby pressuring and tempting scientists to pursue short-term publications through RM (Franzen et al., 2007). The cultural dimension of *Indulgence versus Restraint* also appears some traits related to RM. For example, indulgence, characterized by a propensity to prioritize personal enjoyment over adherence to strict social norms and ethical standards (G. Hofstede et al., 2010, p. 281), may contribute to RM among researchers due to a potential lack of emphasis on ethical considerations and a greater focus on immediate gratification. Its absence in the explanation of cultural factors affecting RM can probably be attributed to its status as a "truly new" concept, available for less than fifteen years (G. Hofstede et al., 2010, p. 281). The relationships between the cultural factors reflected by these two dimensions and RM need to be further tested in future research.

Furthermore, the cultural dimension theory is compatible with other cultural theories in the studies of cultural factors affecting RM. For example, some international students or foreign researchers are likely to commit RM due to the cultural conflict between their home country and host country leading to their divergent awareness of RM, as suggested by Sellin's (1938) theory. As indicated by collectivist culture, researchers who have grown up in collectivist national cultures are more likely to be influenced by and loyal to societal values cultivated in their home country. If

some research behaviours are not accepted as a violation of ethical behaviours in their home countries, it is challenging for them to perceive these behaviours as RM in host countries.

The research findings demonstrated that cultural factors of RM were potentially interrelated with the other three typologies of RM factors that emerged from the preliminary literature review (Table 1), namely individual factors, organizational factors, and professional factors. For instance, '*strong authority hierarchy*' can contribute to '*personal characters*' related to pursuing research success through RM, because Tijdink et al. (2016) once stated that,

> *Alternatively, more pronounced personality traits may be more strongly associated with research misbehaviour in case of high academic position (moderation) due to researchers in top positions being more focused towards success in research.*

Thus, the research team decided to revisit and reanalyse the selected articles from the preliminary review and the main review together to identify evidence supporting underlying relationships. The reanalysis revealed the potential relationships between cultural factors of RM and other typologies of RM factors as shown in Figure 3.

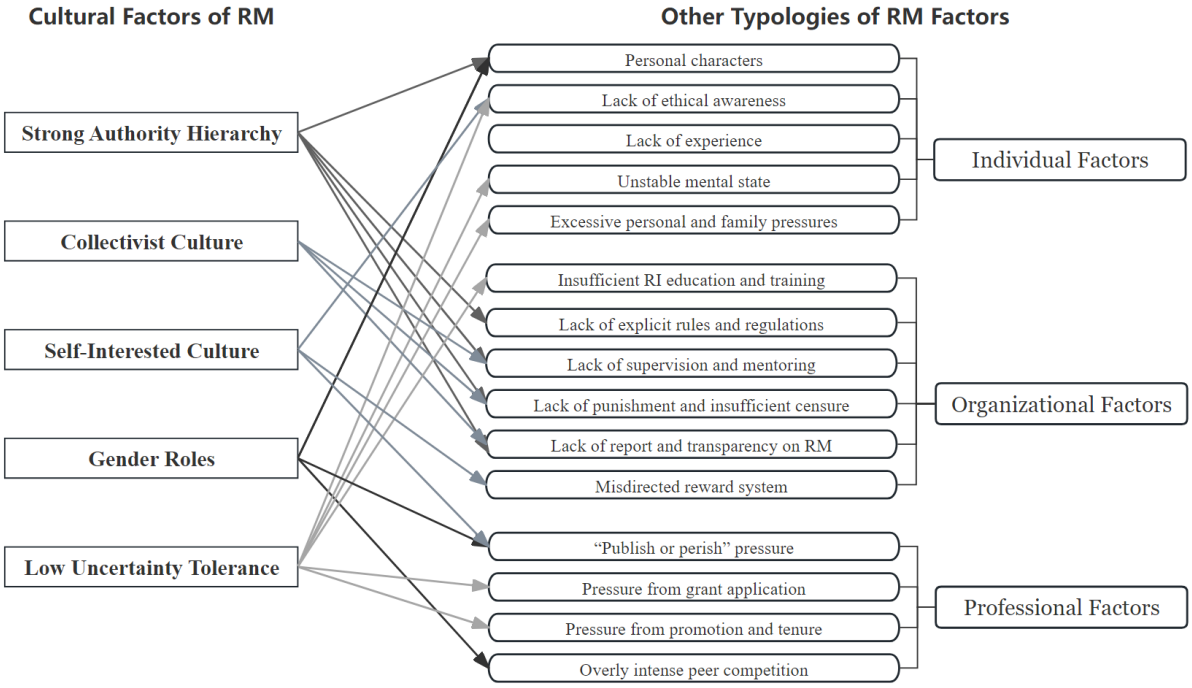


Figure. 3. The relationships between cultural factors and other typologies of factors

As depicted in Fig. 3, strong authority hierarchy and collectivist culture are two principal cultural factors that significantly contribute to the organizational factors of RM. Both cultural factors exhibit a predisposition for deference, with strong authority hierarchy emphasizing deference to authority figures, and collectivist culture emphasizing deference to affiliated institutions or group. A self-interested culture can lead to the misdirected reward system and the '*publish or perish*' pressure in the academic community due to its emphasis on individual gain, recognition, and advancement. This contributes to conditions conducive to RM as individuals might prioritize personal gain over ethical conduct. Additionally, scientists in societies characterized by gender roles, particularly those emphasizing masculinity, may develop certain personal characters such as '*higher aggression, competitiveness, status-seeking, and risk-taking*' (Fanelli et al., 2015, p. 2). These personality traits might make scientists more susceptible to pursuing academic success through RM and can foster an academic environment with overly intense peer competition.

Moreover, the cultural factors of low uncertainty tolerance have instilled in scientists a mindset that resists innovation and seeks security in research, leading to a propensity for resorting to RM to alleviate this unstable mental state.

Therefore, the present study reveals that the cultural factors can contribute to RM by either leading to or combining with other typologies of RM factors, which is in line with Davis's (2003) perspective. However, different from Davis's (2003) model of RM factors, which portrays different RM factors as parallel and overlapping, this systematic review suggests that cultural factors can serve as motivational factors to the emergence of other typologies of RM factors, thereby establishing common '*social origin*' among them (Horbach et al., 2019, p. 416). Thus, a revised conceptual model on the relationships between RM and its factors is formulated and presented in Figure 4. This model illustrates that cultural factors play an indirect role in RM through taking effects on the emergence (represented by arrows to the left) as well as the influencing process (represented by arrows to down) of other RM factors.

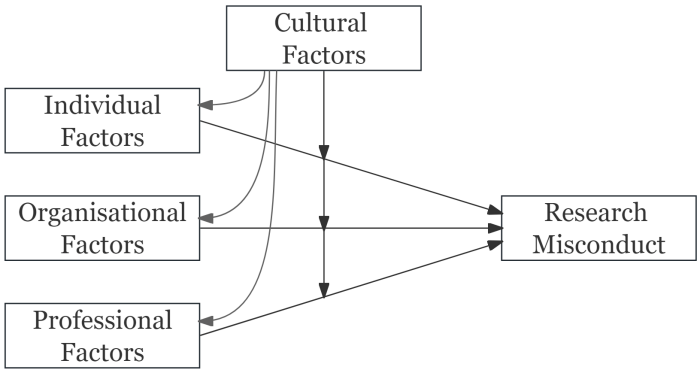


Figure. 4. Revised conceptual model of RM factors

Several implications might be drawn from research findings for both theory and practice. First, the present study contributes to research by shifting the focus on the factors of RM from the tripartite narrative perspective (i.e., individual, organizational, and professional) (Sovacool, 2008) to cultural narrative. Moreover, G. Hofstede's cultural dimension theory is visibly useful for understanding the cultural factors of RM, with four out of six dimensions approved as valid. The five main cultural factors of RM that emerged from the present study can serve as a starting point for future studies to explore the cultural impact on RM. Second, the outcomes of the present study reveal that cultural factors are not influencing RM directly. Instead, they appear to interact with other typologies of RM factors (i.e., individual, organizational, and professional factors) in more complex ways. The revised conceptual model of RM factors provides a map to understand the influencing routine from cultural factors as roots to RM behaviours as destinations. Third, the research findings can help interest groups navigate the complex landscape of RM and formulate RM-preventing strategies based on specific cultural factors and their influencing mechanism. Specifically, RI education measures should pay more attention to foreign researchers and students to help them understand the differences in RI principles between their home cultures and the host cultures. As well, culture reforms at both the national and institutional levels, aimed at addressing the cultural influence on RM issues, should be tailored to specifically target the cultural factors affecting RM.

Conclusion

The present study reports on a systematic review, aiming to investigate the cultural factors of RM and conceptualize their influencing relationships with RM. G. Hofstede's cultural dimension theory was adopted as a theoretical lens to demonstrate that strong authority hierarchy, collectivist culture, self-interested culture, gender roles, and low uncertainty tolerance play a role in RM to

varying degrees. Cultural factors contribute to the connection among other typologies of RM factors and make them no longer separate from each other, thus taking effects together to lead to RM. This systematic review is the first step for theory development, and it can be a solid basis for more empirical studies. To our knowledge, this is one of the earliest studies to construct a comprehensive map of cultural factors affecting RM. Overall, the research findings provide useful insights for interest groups, such as government officers, research institution managers, RI educationists, and information professionals, to develop targeted strategies to mitigate the negative effect of cultural factors on RM. Further, as experts in providing, organizing, locating, and evaluating information, LIS professionals can take on the responsibility of supporting RI and preventing RM guided by the cultural-factor map of RM in the changing academic landscape.

While this study contributes to both theory and practice, certain limitations warrant acknowledgment and should guide future research. First, due to word constraints, the study did not present quite adequate and abundant evidence supporting each code and relationship identified in the thematic analysis. Additionally, the existing literature on the cultural factors influencing RM is limited, resulting in insufficiently robust discussions on some factors. For instance, the exploration of gender tends to rely on stereotypical distinctions between "male" and "female" traits, despite these definitions being drawn from prior studies. Future research should aim for a more nuanced and complex discussion of gender dynamics. Second, while the findings offer plausible explanations for the influence of cultural factors on RM, rigorous empirical studies are required to validate the relationships between RM and specific cultural factors. Particularly, subsequent empirical studies could prioritize the examination of the effects of *Long- versus Short-Term Orientation* and *Indulgence versus Restraint*, which have not been adequately supported by existing research but demonstrate characteristics that impact RM. Third, given that an overreliance on culture as the sole factor of RM is undesirable, it is advisable to study cultural influences on RM in conjunction with other typologies of factors. Especially, there is a need for more cross-cultural studies to compare and contrast the influence of diverse cultural contexts on RM. Fourth, while some measures have been proposed to mitigate the negative impact of culture on RM, it is important to note that cultural factors are not easy to change due to their intricate and persistent nature. As such, future research and practice should focus on developing effective strategies to address the cultural factors of RM issues.

Acknowledgements

This work was jointly supported by the Ministry of Education in China and Beijing Wanfang Data Co., LTD [grant number: 220905469022130] as well as Peng Feizhang Library Science Development Fund Project of Wuhan University. We would like to thank the editors and anonymous reviewers for their comments and suggestions.

About the authors

Wei Feng is research master student at the School of Information Management, Wuhan University, China. She received her bachelor's degree from Wuhan University and her research interests are in human-AI interaction and research integrity. She can be contacted at fengwei211@whu.edu.cn

Lihong Zhou is a professor and the associate dean of the School of Information Management, Wuhan University, China. He obtained his PhD in information studies from the Information School, the University of Sheffield. His research focuses on library services, digital scholarship, and interorganizational information sharing. He can be contacted at l.zhou@whu.edu.cn

Junmin Xiao is a lecturer at the School of Chinese Language and Literature, Central China Normal University, China. She obtained her PhD from the School of Education, the University of Sheffield.

Her research focuses on language education and emergency language. She can be contacted at cathyxiaojm2019@163.com.

References

Afsar, B., & Masood, M. (2018). Transformational leadership, creative self-efficacy, trust in supervisor, uncertainty avoidance, and innovative work behaviour of nurses. *The Journal of Applied Behavioural Science*, 54(1), 36–61. https://doi.org/10.1177/0021886317711891

Aubert Bonn, N., & Pinxten, W. (2019). A decade of empirical research on research integrity: What have we (not) looked at? *Journal of Empirical Research on Human Research Ethics*, 14(4), 338–352. https://doi.org/10.1177/1556264619858534

Aubert Bonn, N., & Pinxten, W. (2021). Rethinking success, integrity, and culture in research (part 2)—A multi-actor qualitative study on problems of science. *Research Integrity and Peer Review*, 6(1), 3. https://doi.org/10.1186/s41073-020-00105-z

Auer, N., & Krupar, E. (2001). Mouse click plagiarism: The role of technology in plagiarism and the librarian's role in combating it. *Library Trends*, 49(3), 415–432.

Bagilhole, B. (2002). Challenging equal opportunities: Changing and adapting male hegemony in academia. *British Journal of Sociology of Education*, 23(1), 19–33.

Benjes-Small, C., Ackerman, E., & Tapp, K. (2008). Exploring the librarian's role in promoting academic integrity on campus. In T. Valko & B. Sietz (Eds.), *LOEX Conference Proceedings 2006* (pp. 123–125). https://commons.emich.edu/loexconf2006/9

Bennett, D. M., & Taylor, D. M. (2003). Unethical practices in authorship of scientific papers. *Emergency Medicine*, 15(3), 263–270. https://doi.org/10.1046/j.1442-2026.2003.00432.x

Bertram Gallant, T. (2015). Leveraging institutional integrity for the betterment of education. In T. A. Bretag (Ed.), *Handbook of Academic Integrity* (pp. 1–12). Springer. https://doi.org/10.1007/978-981-287-079-7_52-1

Bonito, A. J., Titus, S. L., Greene, A. M., Amoozegar, J., Eicheldinger, C., & Wright, D. E. (2012). Preparing whistleblowers for reporting research misconduct. *Accountability in Research*, 19(5), 308–328. https://doi.org/10.1080/08989621.2012.718683

Bornmann, L. (2013). Research misconduct—Definitions, manifestations and extent. *Publications*, 1(3), Article 3. https://doi.org/10.3390/publications1030087

Buljan, I., Barać, L., & Marušić, A. (2018). How researchers perceive research misconduct in biomedicine and how they would prevent it: A qualitative study in a small scientific community. *Accountability in Research*, 25(4), 220–238. https://doi.org/10.1080/08989621.2018.1463162

Buzzelli, D. E. (1993). The Definition of Misconduct in Science: A View from NSF. *Science*, 259(5095), 584–648. https://doi.org/10.1126/science.8430300

Cairns, A. C., Linville, C., Garcia, T., Bridges, B., Tanona, S., Herington, J., & Laverty, J. T. (2021). A phenomenographic study of scientists' beliefs about the causes of scientists' research misconduct. *Research Ethics*, 17(4), 501–521. https://doi.org/10.1177/17470161211042658

Caravello, P. S. (2008). The Literature on Academic Integrity and Graduate Students: Issues, Solutions, and the Case for a Librarian Role. *Public Services Quarterly*, 3(3–4), 141–171. https://doi.org/10.1080/15228950802110502

Casadevall, A., & Fang, F. C. (2012). Reforming science: Methodological and cultural reforms. *Infection and Immunity*, 80(3), 891–896. https://doi.org/10.1128/iai.06183-11

Chapman, D. W., & Lindner, S. (2016). Degrees of integrity: The threat of corruption in higher education. *Studies in Higher Education*, 41(2), 247–268. https://doi.org/10.1080/03075079.2014.927854

Chapman, K. J., & Lupton, R. A. (2004). Academic dishonesty in a global educational market: A comparison of Hong Kong and American university business students. *International Journal of Educational Management*, 18(7), 425–435. https://doi.org/10.1108/09513540410563130

Committee on Responsible Science, Committee on Science, Engineering, Medicine, and Public Policy, Policy and Global Affairs, & National Academies of Sciences, Engineering, and Medicine. (2017). Understanding the causes. In *Fostering Integrity in Research* (pp. 91–104). National Academies Press.

Coseglia, Z., Raad, A., Handron, C., Dowd, L., Irwin, J., & Thomas, K. (2023). *Understanding and Shaping Organisational Culture to Disrupt the Cycle of Misconduct*. https://globalinvestigationsreview.com/guide/the-guide-compliance/second-edition/article/understanding-and-shaping-organisational-culture-disrupt-the-cycle-of-misconduct

Cyranoski, D. (2017). China targets fake peer reviews. *Nature*.

Dannhoferová, J., Foltýnek, T., Dlabolová, D., & Fishman, T. (2022). Cutting the vicious circle: Addressing the inconsistency in teachers' approaches to academic integrity breaches. *Journal on Efficiency and Responsibility in Education and Science*, 15(1), 10–22. https://doi.org/10.7160/eriesj.2022.150102

Davis, M. S. (2003). The role of culture in research misconduct. *Accountability in Research*, 10(3), 189–201. https://doi.org/10.1080/714906092

Davis, M. S., Riske-Morris, M., & Diaz, S. (2007). Causal factors implicated in research misconduct: Evidence from ORI case files. *Science and Engineering Ethics*, 13(4), 395–414. https://doi.org/10.1007/s11948-007-9045-2

DuBois, J. M., Anderson, E. E., Chibnall, J., Carroll, K., Gibb, T., Ogbuka, C., & Rubbelke, T. (2013). Understanding research misconduct: A comparative analysis of 120 cases of professional wrongdoing. *Accountability in Research*, 20(5–6), 320–338. https://doi.org/10.1080/08989621.2013.822248

Durham, W. H. (1976). The adaptive significance of cultural behaviour. *Human Ecology*, 4(2), 89–121. https://doi.org/10.1007/BF01531215

Eckstein, H. (1997). Social Science as Cultural Science, Rational Choice as Metaphysics. In *Culture Matters*. Routledge.

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, 114(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS ONE*, 10(6), e0127556. https://doi.org/10.1371/journal.pone.0127556

Fanelli, D., Schleicher, M., Fang, F. C., Casadevall, A., & Bik, E. M. (2022). Do individual and institutional predictors of misconduct vary by country? Results of a matched-control analysis of problematic image duplications. *PLoS ONE*, 17(3), e0255334. https://doi.org/10.1371/journal.pone.0255334

Fang, F. C., Bennett, J. W., & Casadevall, A. (2013). Males are overrepresented among life science researchers committing scientific misconduct. *mBio*, 4(1), e00640-12. https://doi.org/10.1128/mbio.00640-12

Farahat, A. (2022). Elements of academic integrity in a cross-cultural middle eastern educational system: Saudi Arabia, Egypt, and Jordan case study. *International Journal for Educational Integrity*, 18(1), Article 9. https://doi.org/10.1007/s40979-021-00095-5

Faria, R. (2015). Scientific misconduct: How organizational culture plays its part. *Tijdschrift over Cultuur & Criminaliteit*, 5(1), 38–54. https://doi.org/10.5553/TCC/221195072015005001003

Feng, W., Zhou, L., & Xiao, J. (2024). Beyond the traditional: Extending academic libraries' roles in research integrity based on the causes of research misconduct. *The Journal of Academic Librarianship*, 50(3), 102877. https://doi.org/10.1016/j.acalib.2024.102877

Fierz, K., Gennaro, S., Dierickx, K., Van Achterberg, T., Morin, K. H., & De Geest, S. (2014). Scientific misconduct: Also, an issue in nursing science? *Journal of Nursing Scholarship*, 46(4), 271–280. https://doi.org/10.1111/jnu.12082

Fonseca, P., Olabarrieta-Landa, L., Panyavin, I., Jimenez, X., Arelis, A., Barajas, B., Agudelo, Y., & Lasprilla, J. (2016). Perceived ethical misconduct: A survey of neuropsychology professionals in Mexico. *International Journal of Psychological Research*, 9(1), 64–71. https://doi.org/10.21500/20112084.2101

Fotaki, M. (2013). No woman is like a man (in academia): The masculine symbolic order and the unwanted female body. *Organisation Studies*, 34(9), 1251–1275. https://doi.org/10.1177/0170840613483658

Franzen, M., Rödder, S., & Weingart, P. (2007). Fraud: Causes and culprits as perceived by science and the media. Institutional changes, rather than individual motivations, encourage misconduct. *EMBO Reports*, 8(1), 3–7. https://doi.org/10.1038/sj.embor.7400884

Ghias, K., Lakho, G. R., Asim, H., Azam, I. S., & Saeed, S. A. (2014). Self-reported attitudes and behaviours of medical students in Pakistan regarding academic misconduct: A cross-sectional study. *BMC Medical Ethics*, 15(1), 43. https://doi.org/10.1186/1472-6939-15-43

Gilbert, F. J., & Denison, A. R. (2003). Research misconduct. *Clinical Radiology*, 58(7), 499–504. https://doi.org/10.1016/S0009-9260(03)00176-4

Gillespie, I. E. (1991). Fraud and misconduct in medical research: Causes, investigation and prevention. *Journal of the Royal College of Physicians of London*, 25(2), 89–94.

Golden, J., Mazzotta, C. M., & Zittel-Barr, K. (2023). Systemic obstacles to addressing research misconduct in higher education: A case study. *Journal of Academic Ethics*, 21(1), 71–82. https://doi.org/10.1007/s10805-021-09438-w

Gong, Y. (2005). China science foundation takes action against 60 grantees. *Science*, 309(5742), 1798–1799. https://doi.org/10.1126/science.309.5742.1798a

Gong, Z., Van Swol, L., Xu, Z., Yin, K., Zhang, N., Gul Gilal, F., & Li, X. (2019). High-power distance is not always bad: Ethical leadership results in feedback seeking. *Frontiers in Psychology*, 10, 2137. https://doi.org/10.3389/fpsyg.2019.02137

Gopalakrishna, G., Riet, G. ter, Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLoS ONE*, 17(2), e0263023. https://doi.org/10.1371/journal.pone.0263023

Graham, K. A., Dust, S. B., & Ziegert, J. C. (2018). Supervisor-employee power distance incompatibility, gender similarity, and relationship conflict: A test of interpersonal interaction theory. *Journal of Applied Psychology*, 103(3), 334–346. https://doi.org/10.1037/apl0000265

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Gregson, R., & Kell, P. (2009). An academic problem: An online solution to a clash of cultures. In G. Vogl & P. Kell (Eds.), *Education in the Asia Pacific: Challenges for the future* (pp. 150–164). Cambridge Scholars Publishing.

Gunsalus, C. K. (2019). Make reports of research misconduct public. *Nature*, 570(7759), 7–7. https://doi.org/10.1038/d41586-019-01728-z

Haven, T., & van Woudenberg, R. (2021). Explanations of research misconduct, and how they hang together. *Journal for General Philosophy of Science*, 52(4), 543–561. https://doi.org/10.1007/s10838-021-09555-5

Herr, M. (2019). Responding to research misconduct: A primer for LIS professionals. *Science & Technology Libraries*, 38(3), 272–287. https://doi.org/10.1080/0194262X.2019.1644268

Hofstede, G. (1984). *Culture's Consequences: International Differences in Work-Related Values*. SAGE.

Hofstede, G. (1986). Cultural differences in teaching and learning. *International Journal of Intercultural Relations*, 10(3), 301–320. https://doi.org/10.1016/0147-1767(86)90015-5

Hofstede, G. (1994). The business of international business is culture. *International Business Review*, 3(1), 1–14. https://doi.org/10.1016/0969-5931(94)90011-6

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, 2(1). https://doi.org/10.9707/2307-0919.1014

Hofstede, G., & Bond, M. H. (1984). Hofstede's Culture Dimensions: An independent validation Using Rokeach's Value Survey. *Journal of Cross-Cultural Psychology*, 15(4), 417–433. https://doi.org/10.1177/0022002184015004003

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and Organizations: Software of the Mind* (3rd ed). McGraw-Hill.

Hofstede, G. J. (2000). Organizational culture: Siren or sea cow? A reply to Dianne Lewis. *Strategic Change*, 9(2), 135–137. https://doi.org/10.1002/(SICI)1099-1697(200003/04)9:2<135:AID-JSC464>3.0.CO;2-E

Hofstede, G. J. (2015). Culture's causes: The next challenge. *Cross Cultural Management*, 22(4), 545–569. https://doi.org/10.1108/CCM-03-2015-0040

Hofstede, G., & Minkov, M. (2010). Long- versus short-term orientation: new perspectives. *Asia Pacific Business Review*, 16(4), 493–504. https://doi.org/10.1080/13602381003637609

Hong, Q. N., & Pluye, P. (2018). Systematic reviews: A brief historical overview. *Education for Information*, 34(4), 261–276. https://doi.org/10.3233/EFI-180219

Horbach, S. P. J. M., Breit, E., Halffman, W., & Mamelund, S.-E. (2020). On the willingness to report and the consequences of reporting research misconduct: The role of power relations. *Science and Engineering Ethics*, 26(3), 1595–1623. https://doi.org/10.1007/s11948-020-00202-8

Horbach, S. P. J. M., Breit, E., & Mamelund, S.-E. (2019). Organisational responses to alleged scientific misconduct: Sensemaking, sensegiving, and sensehiding. *Science and Public Policy*, 46(3), 415–429. https://doi.org/10.1093/scipol/scy068

Horbach, S. P. J. M., & Halffman, W. (2019). The extent and causes of academic text recycling or 'self-plagiarism'. *Research Policy*, 48(2), 492–502. https://doi.org/10.1016/j.respol.2017.09.004

Hu, G., & Lei, J. (2012). Investigating Chinese university students' knowledge of and attitudes toward plagiarism from an integrated perspective. *Language Learning*, 62(3), 813–850. https://doi.org/10.1111/j.1467-9922.2011.00650.x

Jam, F. A., Singh, S. K. G., Ng, B.-K., & Aziz, N. (2018). The interactive effect of uncertainty avoidance cultural values and leadership styles on open service innovation: A look at Malaysian healthcare sector. *International Journal of Business and Administrative Studies*, 4(5). https://doi.org/10.20469/ijbas.4.10003-5

James, M. X., Miller, G. J., & Wyckoff, T. W. (2019). Comprehending the cultural causes of English writing plagiarism in Chinese students at a western-style university. *Journal of Business Ethics*, 154(3), 631–642. https://doi.org/10.1007/s10551-017-3441-6

Jesilow, P., Geis, G., Pontell, H., & Song, J. H. L. (1992). Culture conflict revisited: Fraud by Vietnamese physicians in the United States. *International Migration*, 30(2), 201–224. https://doi.org/10.1111/j.1468-2435.1992.tb00693.x

Kaasa, A. (2016). Culture as a Possible Factor of Innovation: Evidence from the European Union and Neighboring Countries. In C. Braedel-Kühner & A. Müller (Eds.), *Re-thinking Diversity: Multiple Approaches in Theory, Media, Communities, and Managerial Practice* (pp. 83–107). Springer Fachmedien. https://doi.org/10.1007/978-3-658-11502-9_5

Kaasa, A., & Vadi, M. (2010). How does culture contribute to innovation? Evidence from European countries. *Economics of Innovation and New Technology*, 19(7), 583–604. https://doi.org/10.1080/10438590902987222

Kaatz, A., Vogelman, P. N., & Carnes, M. (2013). Are men more likely than women to commit scientific misconduct? Maybe, maybe not. *mBio*, 4(2), 10.1128/mbio.00156-13. https://doi.org/10.1128/mbio.00156-13

Kasler, J., Zysberg, L., & Gal, R. (2021). Culture, collectivism-individualism and college student plagiarism. *Ethics & Behaviour*, 31(7), 488–497. https://doi.org/10.1080/10508422.2020.1812396

Krstić, S. B. (2015). Research integrity practices from the perspective of early-career researchers. *Science and Engineering Ethics*, 21(5), 1181–1196. https://doi.org/10.1007/s11948-014-9607-z

Kumar, M. N. (2010). A Theoretical Comparison of the Models of Prevention of Research Misconduct. *Accountability in Research*, 17(2), 51–66. https://doi.org/10.1080/08989621003641132

Lee, C., Huang, G.-H., & Ashford, S. J. (2018). Job insecurity and the changing workplace: Recent developments and the future trends in job insecurity research. *Annual Review of Organizational Psychology and Organizational Behaviour*, 5(1), 335–359. https://doi.org/10.1146/annurev-orgpsych-032117-104651

Li, D., & Cornelis, G. (2021). Differing perceptions concerning research misconduct between China and Flanders: A qualitative study. *Accountability in Research*, 28(2), 63–94. https://doi.org/10.1080/08989621.2020.1802586

Lonner, W. J., & Adamopoulos, J. (1997). Culture as antecedent to behaviour. In J. W. Berry, Y. H. Poortinga, & J. Pandey, *Handbook of Cross-cultural Psychology: Theory and method* (pp. 42–78). John Berry.

Madray, A. (2013). Non-Western students in Western universities: Bridging the plagiarism divide. In L. Chow, C. Fuchs, J. Kucsma, C. N. Wajda, & S. Sajonas (Eds.), *The global librarian* (pp. 207–223). Metropolitan New York Library Council (METRO) and Greater New York Metropolitan Area Chapter/Association of College and Research Libraries, Inc.

Mardani, A., Nakhoda, M., & Gooshki, E. (2020). Relationship among factors affecting research misconduct in medical sciences in Iran. *Accountability in Research-Policies and Quality Assurance*, 27(7), 417–443. https://doi.org/10.1080/08989621.2020.1766977

Martin, D. E. (2012). Culture and unethical conduct: Understanding the impact of individualism and collectivism on actual plagiarism. *Management Learning*, 43(3), 261–273. https://doi.org/10.1177/1350507611428119

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature*, 435(7043), Article 7043. https://doi.org/10.1038/435737a

Mejlgaard, N., Bouter, L. M., Gaskell, G., Kavouras, P., Allum, N., Bendtsen, A.-K., Charitidis, C. A., Claesen, N., Dierickx, K., Domaradzka, A., Reyes Elizondo, A., Foeger, N., Hiney, M., Kaltenbrunner, W., Labib, K., Marušić, A., Sørensen, M. P., Ravn, T., Ščepanović, R., … Veltri, G. A. (2020). Research integrity: Nine ways to move from talk to walk. *Nature*, 586(7829), 358–360. https://doi.org/10.1038/d41586-020-02847-8

Merriam-Webster. (n.d.). *Definition of CULTURE*. Retrieved 4 January 2024, from https://www.merriam-webster.com/dictionary/culture

Merton, R. K. (1938). Social Structure and Anomie. *American Sociological Review*, 3(5), 672. https://doi.org/10.2307/2084686

Michel, M. C. (2010). China fights fraud with tough tactics and integrity training. *Nature*, 463(7283), Article 7283. https://doi.org/10.1038/463877c

Mirshekary, S., & Lawrence, A. D. K. (2009). Academic and business ethical misconduct and cultural values: A cross national comparison. *Journal of Academic Ethics*, 7(3), 141–157. https://doi.org/10.1007/s10805-009-9093-0

Nelson, A., & Lubchenco, J. (2022). Strengthening scientific integrity. *Science*, 375(6578), 247–247. https://doi.org/10.1126/science.abo0036

Olesen, A. P., Amin, L., & Mahadi, Z. (2017). Malaysian researchers talk about the influence of culture on research misconduct in higher learning institutions. *Accountability in Research*, 24(8), 469–482. https://doi.org/10.1080/08989621.2017.1399358

O'Sullivan, M. W., & Guo, L. (2011). Critical thinking and Chinese international students: An east-west dialogue. *Journal of Contemporary Issues in Education*, 5(2). https://doi.org/10.20355/C5NK5Z

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. https://doi.org/10.1136/bmj.n71

Parnther, C. (2022). International students and academic misconduct: Considering culture, community, and context. *Journal of College and Character*, 23(1), 60–75. https://doi.org/10.1080/2194587X.2021.2017978

Peltokorpi, V. (2019). Abusive supervision and emotional exhaustion: The moderating role of power distance orientation and the mediating role of interaction avoidance. *Asia Pacific Journal of Human Resources*, 57(3), 251–275. https://doi.org/10.1111/1744-7941.12188

Petticrew, M., Rehfuess, E., Noyes, J., Higgins, J. P. T., Mayhew, A., Pantoja, T., Shemilt, I., & Sowden, A. (2013). Synthesizing evidence on complex interventions: How meta-analytical, qualitative, and mixed-method approaches can contribute. *Journal of Clinical Epidemiology*, 66(11), 1230–1243. https://doi.org/10.1016/j.jclinepi.2013.06.005

Putra, I. E., Jazilah, N. I., Adishesa, M. S., Al Uyun, D., & Wiratraman, H. P. (2023). Denying the accusation of plagiarism: Power relations at play in dictating plagiarism as academic misconduct. *Higher Education*, 85(5), 979–997. https://doi.org/10.1007/s10734-022-00875-z

Rehman, R., & Waheed, A. (2014). Ethical perception of university students about academic dishonesty in Pakistan: Identification of student's dishonest acts. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2014.1280

Retraction Watch. (2016, September 13). *Diederik Stapel*. https://retractionwatch.com/category/by-author/diederik-stapel/

Rodrigues, F., Gupta, P., Khan, A. P., Chatterjee, T., Sandhu, N. K., & Gupta, L. (2023). The cultural context of plagiarism and research misconduct in the Asian region. *Journal of Korean Medical Science*, 38(12), 1–13. https://doi.org/10.3346/JKMS.2023.38.E88

Roll, L. C., Siu, O., & Li, S. Y. W. (2015). The job insecurity-performance relationship in Germany and China: The buffering effect of uncertainty avoidance. *Psihologia Resurselor Umane*, 13(2), Article 2.

Salter, S. B., Guffey, D. M., & McMillan, J. J. (2001). Truth, consequences and culture: A comparative examination of cheating and attitudes about cheating among U.S. and U.K. students. *Journal of Business Ethics*, 31(1), 37–50. https://doi.org/10.1023/A:1010785106667

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2009). *Research methods for business students* (5th ed). Prentice Hall.

Schiermeier, Q. (2010). German research centre widens misconduct probe. *Nature*. https://doi.org/10.1038/news.2010.671

Schocker, F., Fehrenbach, H., & Schromm, A. B. (2021). Mission impossible? A cultural change to support scientific integrity. *EMBO Reports*, 22(7), e52334. https://doi.org/10.15252/embr.202052334

Schulz, W. (2015). Ethics and the responsible conduct of research in the chemical community: The unique role and challenges of the news media. *Accountability in Research-Policies and Quality Assurance*, 22(6), 384–401. https://doi.org/10.1080/08989621.2015.1047706

Seadle, M. (2017). Research integrity and libraries. In *Research integrity and libraries* (pp. 111–120). De Gruyter Saur. https://doi.org/10.1515/9783110522334-012

Seadle, M. (2023). Looking into the LIS internationalization and information integration from the perspective of iSchools organization (L. Zhou & D. Wang, Trans.). *Documentation, Information & Knowledge*, 40(6), 10–19. https://doi.org/10.13366/j.dik.2023.06.010

Sellin, T. (1938). Culture conflict and crime. *American Journal of Sociology*, 44(1), 97–103. https://doi.org/10.1086/217919

Shi, L. (2006). Cultural backgrounds and textual appropriation. *Language Awareness*, 15(4), 264–282. https://doi.org/10.2167/la406.0

Smith, E., Williamsjones, B., Master, Z., Larivière, V., Sugimoto, C. R., Paul-Hus, A., Shi, M., & Resnik, D. B. (2020). Misconduct and misbehaviour related to authorship disagreements in collaborative science. *Science and Engineering Ethics*, 26(4), 1967–1993. https://doi.org/10.1007/s11948-019-00112-4

Sovacool, B. K. (2008). Exploring scientific misconduct: Isolated individuals, impure institutions, or an inevitable idiom of modern science? *Journal of Bioethical Inquiry*, 5(4), 271–282. https://doi.org/10.1007/s11673-008-9113-6

Staufenbiel, T., & König, C. J. (2010). A model for the effects of job insecurity on performance, turnover intention, and absenteeism. *Journal of Occupational and Organizational Psychology*, 83(1), 101–117. https://doi.org/10.1348/096317908X401912

Stephens, J. M., Romakin, V., & Yukhymenko, M. (2010). Academic motivation and misconduct in two cultures: A comparative analysis of US and Ukrainian undergraduates. *International Journal for Educational Integrity*, 6(1), Article 1. https://doi.org/10.21913/IJEI.v6i1.674

Stern, A. M., Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Financial costs and personal consequences of research misconduct resulting in retracted publications. *eLife*, 3, e02956. https://doi.org/10.7554/eLife.02956

Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. SAGE Publications.

Sun, Y.-C. (2013). Do journal authors plagiarize? Using plagiarism detection software to uncover matching text across disciplines. *Journal of English for Academic Purposes*, 12(4), 264–272. https://doi.org/10.1016/j.jeap.2013.07.002

Tang, L. (2019). Five ways China must cultivate research integrity. *Nature*, 575(7784), 589–591. https://doi.org/10.1038/d41586-019-03613-1

The Journal of Academic Librarianship. (n.d.). *Guide for authors*. https://www.sciencedirect.com/journal/the-journal-of-academic-librarianship/publish/guide-for-authors

The Office of Science and Technology Policy. (2000). *Federal Policy on Research Misconduct*. https://www.govinfo.gov/content/pkg/FR-2000-12-06/pdf/00-30852.pdf

Tijdink, J. K., Bouter, L. M., Veldkamp, C. L. S., Van De Ven, P. M., Wicherts, J. M., & Smulders, Y. M. (2016). Personality traits are associated with research misbehaviour in Dutch scientists: A cross-sectional study. *PLoS ONE*, 11(9), e0163251. https://doi.org/10.1371/journal.pone.0163251

Titus, S. L., Wells, J. A., & Rhoades, L. J. (2008). Repairing research integrity. *Nature*, 453(7198), Article 7198. https://doi.org/10.1038/453980a

Tourish, D., & Craig, R. (2020). Research misconduct in business and management studies: Causes, consequences, and possible remedies. *Journal of Management Inquiry*, 29(2), 174–187. https://doi.org/10.1177/1056492618792621

Valkenburg, G., Dix, G., Tijdink, J., & de Rijcke, S. (2021). Expanding research integrity: A cultural-practice perspective. *Science and Engineering Ethics*, 27(1), 10. https://doi.org/10.1007/s11948-021-00291-z

Velayutham, S., & Perera, M. H. B. (2004). The influence of emotions and culture on accountability and governance. *Corporate Governance: The International Journal of Business in Society*, 4(1), 52–64. https://doi.org/10.1108/14720700410521961

Vučković, D., Peković, S., Blečić, M., & Đoković, R. (2020). Attitudes towards cheating behaviour during assessing students'performance: Student and teacher perspectives. *International Journal for Educational Integrity*, 16(1), Article 13. https://doi.org/10.1007/s40979-020-00065-3

Wang, F., & Zhu, C. (2023). Statistical analysis of research integrity construction in 466 Chinese universities with medical programs. *Humanities and Social Sciences Communications*, 10(1), Article 1. https://doi.org/10.1057/s41599-023-02208-6

Wiedermann, C. (2018). Inaction over retractions of identified fraudulent publications: Ongoing weakness in the system of scientific self-correction. *Accountability in Research-Policies and Quality Assurance*, 25(4), 239–253. https://doi.org/10.1080/08989621.2018.1450143

Williams, S., Tanner, M., Beard, J., & Chacko, J. (2014). Academic misconduct among business students: A comparison of the US and UAE. *Journal of Academic Ethics*, 12(1), 65–73. https://doi.org/10.1007/s10805-013-9200-0

Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics*, 27, Article 41. https://doi.org/10.1007/s11948-021-00314-9

Yi, N., Nemery, B., & Dierickx, K. (2019). How do Chinese universities address research integrity and misconduct? A review of university documents. *Developing World Bioethics*, 19(2), 64–75. https://doi.org/10.1111/dewb.12231

Yu, L., Miao, M., Liu, W., Zhang, B., & Zhang, P. (2021). Scientific misconduct and associated factors: A survey of researchers in three Chinese tertiary hospitals. *Accountability in Research*, 28(2), 95–114. https://doi.org/10.1080/08989621.2020.1809386

Zhou, L., & Nunes, M. B. (2013). Doing qualitative research in Chinese contexts: Lessons learned from conducting interviews in a Chinese healthcare environment. *Library Hi Tech*, 31(3), 419–434. https://doi.org/10.1108/LHT-11-2012-0104

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.

Information Research - Vol. 30 No. iConf (2025)

Researchers' reflections on values, mechanisms, and dynamics of knowledge and information sharing in international academic projects

Dorota Rak, Mahmood Khosrowjerdi, Livhuwani Muthelo, and Bhaskar Mukherjee

DOI: https://doi.org/10.47989/ir30iConf46897

Abstract

Introduction. Knowledge/information sharing is among the vital necessities of effective scholarly collaborations. This study investigates the role of knowledge/information sharing among project members in international academic projects through the lens of Structuration Theory.

Method. The study is based on semi-structured interviews with 28 academicians and researchers from India, Norway, Poland, and South Africa, who have been involved in international academic projects.

Analysis. The transcriptions were imported into F4Analyse software (version 3.4.5) and were coded.

Results. The study reveals three main themes: 1) the significance and value of sharing, 2) mechanisms and platforms for effective communication and sharing, and 3) interpersonal dynamics and challenges in sharing. Each theme highlights critical aspects of the dynamics in international academic collaborations.

Conclusions. The findings underscore the complexities and challenges in these collaborations, such as power dynamics, the potential risks of sharing, and the possibility for exploitation. The study concludes that the process of sharing transcends mere dissemination of information, being deeply rooted in mutual respect, acknowledgment of contributions, and openness. This study contributes to a deeper understanding of the dynamics of information sharing in international academic collaborations.

Introduction

Sustainable Development Goal 17 (SDG 17) emphasizes the importance of partnerships and collaboration for achieving sustainable development goals (United Nations, 2015). Effective information sharing is a critical component of such partnerships, fostering stronger collaboration and enhancing collective action capacities. Songate et al. (2023) highlight the need for intensive knowledge production and transfer in today's competitive global landscape. They stress the significance of information sharing in diversifying and integrating knowledge production actors, highlighting its role in social responsibility amid global disparities. They argue that international academic collaborations, by enabling cross-border knowledge transfer, spur creativity and global innovation. Furthermore, they propose that the internationalization of science enriches knowledge sharing, making it a key driver in advancing science and technology. However, Fosci et al. (2019) investigated the policies and practices of the European research funders on open science and underscored the importance of Incentivising researchers for open sharing of data and information.

Information sharing is a vital component of human activity and plays a crucial role in knowledge management and collaboration within organizations (Ahmad & Huvila, 2019; Ahmad & Karim, 2019; Savolainen, 2019; Widén & Hansen, 2012). Pham and Williamson (2018) explored the interconnectedness of information sharing and collaboration in academic and library staff settings within one Australian and one Vietnamese university. Their research employed an interpretive qualitative approach with data collected from 41 interviews, a focus group, a workshop, and numerous informal conversations with 58 participants. Their analysis revealed that while the Vietnamese university lacked the structural support for collaboration found in the Australian counterpart, this was compensated by the initiatives of the library staff, demonstrating the power of human agency. Trust and technology emerged as supportive elements in promoting effective collaboration and information sharing. The researchers concluded that successful collaboration relies on information sharing and vice versa. They underscored the role of supportive infrastructure, human agency, trust, and technology in fostering collaboration and information sharing.

Sharing information/knowledge can occur not only in organizations with formalized physical structures but also in temporary constructs established for a specific purpose, such as research teams undertaking international scientific collaborations or teams whose collaboration is based on joint works on projects. Information sharing is integral to effective collaboration, and its effectiveness is influenced by various factors, including organizational context and project characteristics (Ren et al, 2019). Effective knowledge sharing in multi-stakeholder collaborations hinges on several factors, including a common understanding and commitment to the project's goals, initial respect and trust among members, and the presence of strong, charismatic leaders (Molina & Yoong, 2003). Trust has been regarded as an important antecedent of information sharing, and any lack of trust can increase the likelihood of miscommunication and withholding of information, which in turn can affect the success of the project and the sustainability of the partnership (Yue et al, 2022). Besides, concerns about confidentiality and limited access to resources can hinder information sharing and could lead to an increased risk of errors and influence work quality (Auschra, 2019). Despite the recognized importance of information sharing, there is a gap in understanding how knowledge/information sharing among project members contributes to the success or failure of international academic projects. This gap is particularly evident in the context of temporary constructs such as research teams in international scientific collaborations. Therefore, our study aims to address this gap by exploring the following overarching questions:

- How can the process and models of knowledge/information sharing among project members influence the success or failure of international academic projects?

- What experiences or reflections can inform the selection of the most effective information sharing model?

To guide this exploration, the study will utilize Giddens' Theory of Structuration (Giddens, 1984). The theory provides a robust framework for studying information sharing among researchers by integrating the interplay of structure and agency. It explains how human behavior is shaped by both individual actions and the social structures that these actions produce and are constrained by. This duality of structure shows that social structures are both the medium and outcome of repeated practices. Researchers' information-sharing behaviours are influenced by institutional structures like academic norms, funding bodies, and policies, as well as individual actions such as data sharing and collaboration. Additionally, the theory addresses the impact of power relations and resource access on researchers' ability and willingness to share information, which is essential for identifying barriers and developing effective strategies.

The study will provide a deeper understanding of the role of information sharing in the success or failure of international academic projects and will potentially inform the development of more effective information sharing models for such collaborations, contributing to the successful realization of SDG 17.

Method

Our study is based on a research project conducted using semi-structured interviews (including 12 questions on international academic collaborations) with 28 academicians and researchers who had prior involvement in at least one international academic project. The participants were purposefully selected from four diverse countries: India, Norway, Poland, and South Africa. These countries span a broad spectrum of geographical regions, cultural contexts, and economic conditions, encompassing Eastern European, Nordic, African, and South Asian cultures. This rich cultural mix, combined with varying levels of English proficiency, amplifies the diversity of the sample. This comprehensive representation enables an in-depth exploration of academic collaborations in different contexts. The features of participants are drawn in Table 1.

Interviewee code	Discipline	Working place	Gender	Age	Academic Rank
Indian sample (8)					
IND-HUM1-Univ-F-P	HUM	Univ	F	51-60	P
IND-HUM2-Univ-M-A	HUM	Univ	M	61+	A
IND-MET1-RC-M-P	MET	RC	M	51-60	AP
IND-MET2-Univ-M-P	MET	Univ	F	61+	P
IND-NAS1-Univ-F-P	NAS	Univ	F	61+	P
IND-NAS2-Univ-M-P	NAS	Univ	M	51-60	P
IND-ET1-Univ-M-P	ET	Univ	M	51-60	P
IND-ET2-Univ-M-P	ET	Univ	M	51-60	P
Norwegian sample (8)					
NOR-HUM1-Univ-F-AP	HUM	Univ	F	41-50	AP
NOR-HUM2-Univ-M-P	HUM	Univ	M	61+	P
NOR-MET1-UnivCol-F-AP	MET	UnivCol	F	41-50	AP
NOR-MET2-Univ-M-AP	MET	Univ	M	41-50	AP

NOR-NAS1-UnivCol-M-P	NAS	UnivCol	M	51-60	P
NOR-NAS2-UnivCol-F-AP	NAS	UnivCol	F	41-50	AP
NOR-ET1-Univ-F-P	ET	Univ	F	51-60	P
NOR-ET2-UnivCol-M-ASP	ET	UnivCol	M	50-60	ASP
Polish sample (6)					
POL-HUM1-Univ-F-AP	HUM	Univ	F	41-50	AP
POL-HUM2-Univ-F-ASP	HUM	Univ	F	51-60	ASP
POL-NAS1-Univ-M-AP	NAS	Univ	M	51-60	AP
POL-NAS2-Univ-F-AP	NAS	Univ	F	41-50	AP
POL-ET1-Univ-M-ASP	ET	Univ	M	30-40	ASP
POL-ET2-Univ-M-P	ET	Univ	M	41-50	P
South African sample (6)					
SA-HUM1-UNIV-M-P	HUM	Univ	M	30-40	P
SA-HUM2-UNIV-M-P	HUM	Univ	M	51-60	P
SA-MET1 -UNIV-M-ASP	MET	Univ	M	30-40	ASP
SA-MET2-UNIV-F-P	MET	Univ	F	51-60	P
SA-NAS1-UNIV-F-ASP	NAS	Univ	F	30-40	ASP
SA-NAS2-UNIV-F-ASP	NAS	Univ	F	41-50	ASP

Note for table 1.
Disciplinary codes: HUM: Arts, humanities, and social sciences, MET: Medical and health sciences, NAS: Natural and agricultural sciences (NAS), and ET: Engineering and technology
Gender codes: F (female), M (Male)
Working place codes: Univ (University), UnivCol (University college), RC (Research institute/centre)
Academic rank codes: ASP (Assistant professor), AP (Associate professor), P (professor), A (Academic Advisor)

Table 1. The features of interviewed participants (n=28)

For this study, the data on questions regarding knowledge and information sharing during international academic collaborations were analysed. These questions were formulated as follows in the interviews:

> *how could the knowledge/information sharing among project members contribute to the success/failure of the international collaborative projects? Do you have any experience with/reflections on that? Which information sharing model could work better than others?*

The transcriptions were imported into F4Analyse software (version 3.4.5) and were coded. The draft of codes was extracted from the transcription by one author, and the co-authors critically assessed the extracted codes and emerged themes to check the relevance and precision of assigned codes and themes. The reliability and validity of the analyses were sought via three approaches mentioned by Merriam and Grenier (2019, p. 31): 1) triangulation (using multiple

sources of data, this is, cross-country data), 2) variations and diversity (via covering a diverse sample from various disciplines for greater applicability of the results in practice), and 3) deep and rich descriptions (i.e. providing enough description for contextualization of the study and transferability of the findings). The emergent themes/codes were discussed and decided by all authors.

Results

Three main themes emerged from our data that capture the essence of information sharing behavior among researchers in international academic collaborations. These themes, each associated with a set of codes, provide a comprehensive understanding of the dynamics at play in such collaborative environments.

The first theme, ***the significance and value of sharing***, underscores the importance and perceived value of information sharing in these collaborative settings. This theme encompasses codes such as the importance of sharing, valuing shared information and respecting experience/expertise, and willingness to share and trust. These codes collectively highlight the critical role of sharing in enhancing the overall productivity and success of international academic collaborations.

The second theme, ***mechanisms and platforms for effective communication and sharing***, pertains to the mediums and mechanisms through which information is shared. The emergence of codes like effective communications and forums, safe systems/platforms, physical meetings, and documented guidelines/organized information under this theme indicates the need for robust and secure platforms that facilitate efficient communication and sharing of information.

The third theme, ***interpersonal dynamics and challenges in sharing***, explores the interpersonal factors and potential hurdles in the information sharing process. The associated codes include collective action, power tensions, risks of sharing, and the leaders' role. These codes shed light on the complexities and conflicts that might arise in information sharing and underscore the need for effective leadership to navigate these challenges.

Theme 1. The significance and value of sharing

The first theme of the study is centred around the importance of sharing, valuing shared information, and the willingness to share and the role of trust in the information sharing process.

Importance of sharing

Researchers from all four countries have emphasized the benefits of sharing expertise and experiences. Such sharing can introduce new references, resources, and perspectives that might not have been considered otherwise, contributing to a project's success. However, inefficient knowledge exchange could potentially harm project implementation.

> *The benefits [of information sharing], obviously, is that you are sharing the expertise and the experiences of a wide variety of people that may differ and may be able to complement each other.* (SA-NAS2-UNIV-F-ASP).

> *By openly sharing knowledge, members bring in new references, resources, and perspectives that may not have been considered otherwise.* (IND-HUM1-Univ-F-P).

> *I actually think that we have quite good information sharing, and I think that it is important to be, positively inclined.* (NOR-MET1-Univ-F-AP).

> *As the exchange of knowledge and information is not efficient, it can harm the implementation of the project.* (POL-HUM2-Univ-F-ASP).

Valuing shared information, respecting experience or expertise

Our data reveals the importance of attentiveness to shared information and acknowledging each other's thoughts and input. Researchers from Norway and Poland emphasized the importance of valuing the contributions of more experienced team members. An Indian researcher highlighted the role of mutual respect in successful collaboration, and a Norwegian researcher suggested that credit/incentives could be given for sharing information or data.

> *… listening to the dialogue and being very attentive to the information shared is very important… when you work on things together, that you acknowledge each other and acknowledge each other's thoughts, ideas, and input, and not necessarily go directly in and criticize.* (NOR-MET1-Univ-F-AP).

> *If we say that with such a strictly project that, for example, we have to determine, I don't know, the structure of questionnaires, research questions, well, here people with more experience, well, they can certainly give a greater contribution.* (POL-HUM1-Univ-F-AP).

> *… mutual respect is equally important. While one may be ready and willing to share knowledge, it is vital that the other members value and respect the shared information, even if they hold differing opinions. Successful collaboration is not just about sharing but also about acknowledging and appreciating the contributions of others, fostering an environment of intellectual respect and openness.* (IND-HUM1-Univ-F-P).

> *… maybe one should have gotten a little credit [for sharing information/data] in a way.* (NOR-NAS2-UnivCol-F-AP)

Willingness to share and trust

Researchers from Norway, Poland, and India emphasized the importance of motivation and openness in the knowledge-sharing process. It was noted that the level of engagement in knowledge sharing can vary among individuals, ranging from full involvement to minimal participation. Effective collaboration was seen to hinge on the willingness of all members to share their knowledge.

> *People look at it [knowledge sharing] in a very different way, that some are all in, and that others are [at] a minimum [level]* (NOR-NAS2-UnivCol-F-AP).

> *As I said, this very motivation is basically everything, so openness is the most important thing, so it's better to communicate even this negative knowledge than to hide something and later have problems because of it*(POL-ET1-Univ-M-ASP).

> *Knowledge sharing is critical for the success of international academic projects. Effective collaboration depends on the willingness of all members to share their knowledge. My experience shows that projects with open knowledge sharing tend to succeed, while reluctance to share knowledge can hinder progress*(IND-HUM2-Univ-M-P).

Researchers from all four countries highlighted the role of trust in successful sharing. Trust among partners was seen as a critical factor that enables the willing sharing of resources without fears of exploitation. It was also noted that a lack of trust could lead to the independent evolution of projects, despite the existence of a collaborative relationship. Positive experiences involving respect for intellectual property and effective communication were seen as crucial to fostering trust and willingness to share.

> *Knowledge sharing it will lead to a synergistic success. but if there's no trust the two projects' collaborators they might have to actually evolve independently, even though they are using their relationship to grow certain parts* (SA-NAS1-UNIV-F-ASP).

> *… underlying factor for willingness is trust among collaborating partners. When trust is present, partners share resources more willingly without any fear of exploitation* (IND-NAS2-Univ-M-P).

> *I have had positive experiences with my collaborators, who have respected intellectual property and refrained from appropriating my ideas for their own publications* (IND-ET1-Univ-M-P).

> *... as well as trust, effective communication [is] crucial at every stage of projects undoubtedly*(POL-NAS1-Univ-M-AP).

Theme 2. Mechanisms and platforms for effective communication and sharing

The second theme sheds light on the different methods and platforms utilized for information sharing in international academic collaborations.

Physical meetings

Researchers from India, Norway, and Poland highlighted the value of face-to-face meetings for effective communication and understanding. Physical meetings were considered even more beneficial than digital ones, as they provided more context and facilitated better discussions.

> *Scheduling planned face-to-face meetings for information sharing tends to be even more beneficial. These physical meetings facilitate better communication and understanding, thereby contributing significantly to the success of international collaborations*(IND-MET1-RC-M-P).

> *Normally, we have a meeting among the collaborators per week/ month to share our experiences to improve the collaboration* (IND-MET2-Univ-M-P).

> *Physical meetings are much better than digital, because you lose a big part of the context or the information around the thing that is shared when we have digital meetings*(NOR-MET1-Univ-F-AP).

> *Well, and then we share knowledge and also gain knowledge from other areas and also discuss among ourselves. Yes, this is very important, and this is a bit lacking. And if we talk about international cooperation, which is done online, these meetings are dedicated to specific issues and it's more difficult to share this knowledge* (POL-HUM1-Univ-F-AP).

Effective communications and forums

Researchers from South Africa, Norway, and Poland emphasized the importance of open dialogue and organized events like seminars and webinars for knowledge sharing. Projects often involve members attending conferences together to absorb '*condensed doses of information*'. However, the need for effective communication was highlighted, specifically in interpreting data and understanding specific project details.

> *Initially when we started [our project] before COVID-19, we would rotate countries as partners and have mini conferences* (SA-HUM1-UNIV-M-P).

> *Knowledge sharing is very much important because if ever a certain member knows how things can be done, need to share with others so that they can have knowledge of that … it is so important that members of the project organize some webinars, seminars and share their expertise* (SA-MET2-UNIV-F-P).

> *And at some point, [as a project manager,] I get the manuscript describing the field work and I can read that they have been interviewing 400 people, for example, which is a huge amount of work compared to what I thought would be the design for that study. So, it's maybe in a sense it may indicate a lack of information or at least on the details to me. But then again, as a project leader, I don't need to know everything* (NOR-NAS1-Univ-M-P).

> *As well as trust, effective communication [is] crucial at every stage of projects undoubtedly… the exchange of [information about] what did you actually mean there when you wrote that particular passage or what does this data mean? or how to interpret the data in the context of your country? [are important]* (POL-NAS1-Univ-M-AP).

> *When it comes to sharing knowledge, well, certainly it's conferences, and that's why it's worth going to them to hear such condensed doses of information, and it's often the case that if we have a project, it's several members of a research project who go to a conference together* (POL-HUM1-Univ-F-AP).

The role of safe systems/platforms

Researchers from South Africa, India, and Poland suggested the use of secure online solutions like Google Drive or Dropbox for sharing information, however, they are lacking an integrated platform for information sharing. The importance of a centralized information sharing model, where all relevant data is accessible in one place, was also emphasized for maintaining clarity and coordination.

> *We just need to have an online platform, a secure space where you can share that [information that you would like to share with project members*] (SA-MET1-UNIV-M-ASP).

> *In my experience, a centralized information sharing model, where all relevant data is accessible in one place, tends to work better for maintaining clarity and coordination* (IND-ET2-Univ-M-P).

> *I've seen also some people create a Google Drive or a Dropbox where they drop everything that they think will be relevant to that project*(SA-NAS1-UNIV-F-ASP).

> *We don't have a platform; we use solutions all the time…it's just a difficulty that I haven't yet found how to solve because this information sometimes gets lost. We use all the tools, we have Google Docs, on which we do summaries, so that it is easier to get to it, but if these sources of information of those people who participate are so many, well, here obviously there is often a certain chaos of information* (POL-HUM1-Univ-F-AP).

Documented guidelines/organized information

Researchers from India, Poland, and South Africa pointed out the need for efficient organization of information and the necessity of having agreements or values governing the sharing of information. This ensures that shared ideas and information remain confidential among the project members.

> *From my experience, having a clearly written document or record combined with telephonic discussions or explanations can be very effective*(IND-MET1-RC-M-P).

> *There is supposed to be an exchange of knowledge and information, and it is supposed to be organized efficiently so that it works, as it should. This is the basis of the project* (POL-HUM2-Univ-F-ASP).

> *You need to have values to govern the sharing of that information* (SA-HUM1-UNIV-M-P).
>
> *Sharing [is always] very important, but we need to have an agreement that our idea [or shared information] must remain between us [the project members] (*SA-MET1-UNIV-M-ASP).

Theme 3. Interpersonal dynamics and challenges in sharing

Theme 3 highlights the complexities and challenges that arise in the interpersonal dynamics of academic collaborations, particularly in the context of sharing.

Collective 'action'

Researchers from Poland and South Africa emphasized the value of cooperative work and the exchange of knowledge and experiences. A researcher from South Africa preferred the term '*action*' as it implies active participation in the knowledge-sharing process.

> *I have already said - it depends a bit on how the project is conceived. I prefer projects where there is an obvious element of such cooperation and based on the exchange of knowledge, experiences, and also results, where these results are achieved together, jointly*(POL-ET2-Univ-M-P).
>
> *Consequently, I have an outsider's perspective on certain things, and they have an insider's [perspective], and together we create something interesting, right?* (POL-NAS2-Univ-F-AP).
>
> *I can't say, this [information sharing] one can work better than the other, but working together as a group, let me say, [collaborative] actions [are important] ... I prefer [the word] action because if I see you [are] doing, then it is so easy for me to do the same thing* (SA-MET2-UNIV-F-P).

Power tensions

Researchers mentioned the potential for unequal power dynamics within collaborations. Researchers from India, South Africa, and Poland discussed situations where collaborators could use shared ideas to their advantage, particularly if they have access to advanced technology or resources. There were also instances where partners might be interested in collaboration but not willing to invest substantial effort in sharing information effectively.

> *I am aware that there are instances where collaborators might leverage openness and shared ideas to their advantage, particularly when they have access to advanced instrumentation. In such cases, while the initial idea remains crucial, the superior technology at their disposal can accelerate implementation and lead to further recognition. Thus, while the integrity of collaborative exchanges is generally maintained, the availability of high-end equipment can significantly impact the pace and acknowledgment of the research outcomes* (IND-ET1-Univ-M-P).
>
> *Some people are selfish with their knowledge or with information, let's say tools. Some people want to always be the first to try a certain tool and when it works and all that. So, we just have to be open and honest about this process of information sharing* (SA-HUM1-UNIV-M-P).
>
> *There are situations in which partners seem to want to implement the project together, but they want to limit... or maybe I will answer the opposite, that they do not want to limit, but they do not want to invest too much effort in such an exchange* (POL-ET2-Univ-M-P).

> *I have the impression, however, that often Central and Eastern Europe, at least Poland, is treated as a place to obtain data, which will then be beautifully processed to publish a nice article, for example, already without the participation of Polish scientists or with a small participation of Polish scientists. And then I think that this is a failure of the project so that both sides should have something from this cooperation, because it is such, well, for me it is still a colonial approach* (POL-NAS2-Univ-F-AP).

Risks of sharing

Researchers from Norway and South Africa highlighted possible issues, such as the '*free rider problem*' where some partners could exploit the shared information for their own ends without active contributions in the same way. The importance of having values to govern the sharing of information was stressed to prevent such exploitations.

> *There you can get to be a bit so free rider problem, or a bit so command problem, that everyone wants everyone else to share, but one is not sure that everyone will share together* (NOR-NAS2-UnivCol-F-AP).

> *You need to have values to govern the sharing of that information… You can start collaboratively, but somewhere along the line, you will always have those who advance their selfish interests at the cost of the team… if you put it [your data/information] there, some people might exploit it for their own ends or selfish interests before they realize it. Already your data is published by somebody who's a partner coming from [another continent]* (SA-HUM1-UNIV-M-P).

> *I share my idea you take my idea you shot to another person who [is] not part of operation [i.e., project] (*SA-MET1-UNIV-M-ASP).

Leader's role

Researchers from Poland and Norway discussed the importance of a project leader's role in information sharing. Key responsibilities of a project leader include collecting, distributing, and organizing information. A good project leader needs to have not just management skills but also human skills like the ability to involve everyone in discussions and to ensure each person's opinion is heard.

> *One of the main tasks of such a [project] leader is to collect this information, because there is a certain chaos, and this is the difficulty. So, it must be a person who is responsible to collect it, distribute it and organize it* (POL-HUM1-Univ-F-AP).

> *Well, projects, it's kind of a matrix management. [It is successful] If you have a good project leader. [The project leader] has the ability to attract the smartest in the room…as a leader, is to dare to ask people for their opinion directly… I use myself as an example. If I'm an introverted person, I can sit back and listen to the discussion, because that's my mode. But a good project leader will not let me sit quietly all the time. So being a bit specific about everyone getting the opportunity to say something, I think that has to do with management skills… It can be management skills, but also human skills to do, I think. That you choose to ask for exactly the information each person may have. Dig an extra time [and ask]: Is there anyone who has experience with this or wants to do this* (NOR-MET2-Univ-M-AP).

Discussions

The findings from our study underscore the significance and value of knowledge and information sharing in the context of international academic collaborations. Participants (academicians and researchers) across all four countries of Norway, Poland, India, and South Africa highlighted the benefits of sharing expertise and experiences, implying that the dissemination of knowledge within

these collaborative frameworks contributes significantly to the success of a project. The participants highlight that the value of sharing is primarily derived from the infusion of new sources, resources, and perspectives that might not have been considered otherwise. However, they also mention that inefficient knowledge exchange could potentially harm project implementation, emphasizing the need for structured and efficient sharing mechanisms.

Our study revealed the importance of respecting and valuing shared information and the contributions of team members. Attentiveness to shared information and acknowledging each other's thoughts and ideas is reported by participants as a critical aspect of successful collaborations. A Norwegian participant mentioned the idea of giving credit/incentive for shared information or data, and this could be explored further in future research as it might incentivize more open sharing of knowledge/information/data in academic projects. Similar results have been reported in literature (see for example, Fosci et al., 2019) on the role of incentives and their alignments with the whole research system. This is a good ground for supporting initiatives such as the toolbox for recognition and rewards in academic careers developed by Universities Norway (UHR, 2021) in which open sharing of research data has been regarded as a positive criterion for promotions of researchers.

The willingness to share and trust stood out as a crucial determinant of effective knowledge sharing and collaboration. As emphasized by interviewed participants in this study, motivation and openness play a significant role in the knowledge-sharing process. The participants also highlighted trust among partners as a critical factor that enables the willing sharing of resources without fears of exploitation. A lack of trust could lead to the independent evolution of projects, despite the existence of a collaborative relationship. These findings suggest that to facilitate open and effective knowledge sharing among project members, fostering trust is very vital, and they are in accordance with previous findings of Yue et al. (2022), in which, the positive role of trust in sustainable performance, and the negative consequences of lack of trust and withholding information have been highlighted.

The participants also point out to different methods and platforms utilized for information sharing in international academic collaborations. Participants valued face-to-face meetings for effective communication and correct understanding of shared information, considering them even more beneficial than digital ones. The use of online solutions like Google Drive or Dropbox for sharing was also suggested, but the need for a more integrated platform for information sharing was highlighted, emphasizing the importance of a centralized information sharing model where all relevant data is accessible in one place.

The findings of this study exhibit the complexities and challenges that arise in the interpersonal dynamics of academic collaborations. Power tensions were noted, with participants discussing situations where collaborators could use shared ideas to their advantage, particularly if they have access to advanced technology or resources. The ´free rider problem´ where some partners could exploit the shared information for their own ends, was also mentioned by participants that underscores the need for values to govern the sharing of information. The participants emphasized the role of a project leader in collecting, distributing, and organizing information, the need for management skills, and human relations skills.

These findings are in line with findings of Pham and Williamson (2018) on the role of infrastructure, human agency, trust, and technology in information sharing in the organizational context, and the fruitfulness of Structuration Theory in understanding their connections. The Structuration Theory posits that social practices are continuously being made and remade through an ongoing interplay between structures and agency (Giddens, 1984, pp. 1-29). The importance of sharing, as highlighted in our findings, can be seen as an aspect of the structure in Structuration Theory. It's a social practice that has been established as beneficial and necessary in international academic

collaborations. This structure is upheld through the agency of individuals, in this case, researchers and academicians, who actively participate in sharing their expertise and experiences. The structure thus both constrains and enables the agency—it guides the researchers on how they should behave in collaboration (i.e., share knowledge, information, and data) but also provides them the opportunity to add new perspectives and resources to the project. Written agreements can be considered a type of structure that offers rules (the terms of the agreement) and resources (the information shared) for the individuals involved (researchers in international projects). While these agreements serve as a formal structure, they are not strictly inhibiting or confining. Instead, they facilitate information sharing by setting clear expectations, outlining roles, and creating a shared framework for collaboration.

The role of trust in international academic collaborations previously mentioned can also be examined through Structuration Theory. Trust could be viewed as a structural property of these collaborations, a norm that has been established as critical for the willing sharing of resources. Simultaneously, trust is also perpetuated and strengthened (or weakened) through the actions of the individuals involved in the collaboration. Positive experiences and mutual respect can foster trust, while instances of the ´free rider problem´ can erode it.

Our findings relating to the use of different methods and platforms for information sharing can be seen as the modality in the theory, which is the means through which the structure is translated into action. The choice of face-to-face meetings or online solutions like Google Drive or Dropbox are the modalities that the researchers employ to navigate the structure of academic collaborations.

Furthermore, the complexities and challenges highlighted in the findings, such as power dynamics and the risks of sharing, can be seen as the contradictions that could be extracted from the theory. Giddens (1984, p. 14) defines action as the capacity of individuals to intentionally impact their surroundings or alter existing situations. He posits that if an individual loses the ability to make a difference, it ceases to function as an agent:

> *action depends upon the capability of the individual to make a difference to a pre-existing state of affairs or course of events. An agent ceases to be such if he or she loses the capability to make a difference, that is, to exercise some sort of power* (Giddens, 1984, p. 14).

Conclusion

This study underlines the interconnectedness of sharing, trust, and the role of effective leadership in international academic collaborations. It reveals that the process of sharing is not just about disseminating information but is also deeply rooted in mutual respect, acknowledgment of contributions, and the willingness to be open. It also highlights the need for secure and efficient platforms for knowledge sharing and the importance of leaders who can effectively manage and facilitate the process. Besides, it shows the complexities and challenges inherent in these collaborations, such as power dynamics, the risks of sharing, and the potential for exploitation.

Researchers from India, Norway, Poland, and South Africa highlighted the critical role of trust in successful knowledge sharing, and the level of trust and willingness to share were relatively similar across cultures. Indian and South African researchers stressed the necessity of mutual respect and trust for open sharing, while Norwegian and Polish researchers noted that without sufficient trust, the collaborative efforts can break down, leading to parallel but uncoordinated efforts rather than cohesive joint projects. Furthermore, while Indian researchers prioritized mutual respect and acknowledgment, Norwegian researchers suggested incentives, and Polish researchers emphasized respecting contributions from experienced team members. Power dynamics were a concern in all countries, with Indian and South African researchers particularly noting the risk of

exploitation of shared ideas, highlighting the need for clear values and agreements. Additionally, the role of leadership was emphasized, with Polish and Norwegian researchers underscoring the need for project leaders to effectively manage information sharing, navigate power dynamics, and ensure equitable participation.

Future research could explore these dynamics further and work towards developing strategies and mechanisms to mitigate these challenges and enhance the effectiveness and impact of international academic collaborations.

This study provides valuable insights into knowledge and information sharing in international academic collaborations but has some limitations. The qualitative data was gathered through purposive sampling from a small number of researchers (who filled our inclusion criteria), which may introduce selection bias and not fully capture the diversity of experiences in international projects. This limitation should be considered when interpreting and applying the study's findings.

Acknowledgements

The authors would like to thank all participants who participated in the interviews. The valuable comments and suggestions of the reviewers and editors on the early version of the manuscript are acknowledged.

About the authors

Dorota Rak is an Assistant Professor in the Institute of Information Studies at Jagiellonian University, Kraków, Poland. She received her Ph.D. from the Jagiellonian University. Her research interests include Information and Digital Literacy, marketing and communication, and the co-creative processes of innovations in libraries. She can be contacted at dorota.rak@uj.edu.pl.

Mahmood Khosrowjerdi is a Senior Academic Librarian at the University of Inland Norway, Elverum, Norway. He received his Ph.D. from the Oslo Metropolitan University (OsloMet). His research interests focus on cross-cultural studies, collaborations, and health information studies. He can be contacted at mahmood.khosrowjerdi@inn.no.

Livhuwani Muthelo is an Associate Professor in the Department of Nursing Science School of Health Science University of Limpopo South Africa. She received her Ph.D. from the University of Limpopo, and her research interests are in women, occupational and digital health. She can be contacted at Livhuwani.muthelo@ul.ac.za

Bhaskar Mukherjee is a Professor in the Department of Library & Information Science at Banaras Hindu University, Varanasi, India. He received his Ph.D. from Banaras Hindu University on Scholarly Communication in Open Access Journals. His research interests are in bibliometrics, scientometrics, and natural language processing. He can be contacted at bhaskar.mukherjee1@bhu.ac.in.

References

Ahmad, F., & Huvila, I. (2019). Organizational changes, trust and information sharing: an empirical study. Aslib Journal of Information Management, 71(5), 677-692.

Ahmad, F., & Karim, M. (2019). Impacts of knowledge sharing: a review and directions for future research. Journal of workplace learning, 31(3), 207-230.

Auschra, C. (2018). Barriers to the integration of care in inter-organisational settings: a literature review. International journal of integrated care, 18(1).

Fosci, M., Richens, E., & Johnson, R. (2019). Insights into European research funder: policies and practices. SPARC Europe. https://www.ouvrirlascience.fr/wp-content/uploads/2019/10/SPARC-Europe_FundersOS_Report2019_Final.pdf

Giddens, A. (1984). The Constitution of Society: Outline of the Theory of Structuration. University of California Press.

Merriam, S. B., & Grenier, R. S. (Eds.). (2019). Qualitative research in practice: Examples for discussion and analysis. John Wiley & Sons.

Molina, M., & Yoong, P. (2003). Knowledge sharing in a co-opetitive environment: The case of business clusters. Journal of Information & Knowledge Management, 2(04), 321-341.

Pham, H. T., & Williamson, K. (2018). A two-way street: collaboration and information sharing in academia. A theoretically-based, comparative Australian/Vietnamese study. Information Research, Vol. 23 No. 4. https://informationr.net/ir/23-4/isic2018/isic1810.html (accessed September 14, 2024).

Ren, X., Yan, Z., Wang, Z., & He, J. (2019). Inter-project knowledge transfer in project-based organizations: an organizational context perspective. Management Decision, 58(5), 844-863.

Savolainen, R. (2019). Modeling the interplay of information seeking and information sharing: A conceptual analysis. Aslib Journal of Information Management, 71(4), 518-534.

Songate, E. K., Maren, B., Bhandari, N., & Valia, C. (2023). Knowledge Transfer and Management Beyond Borders: An Indo-German Collaboration Perspective. In Reimagining Border in Cross-border Education (pp. 45-73). Routledge.

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development. https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

Universities Norway (UHR) (2021). A toolbox for recognition and rewards in academic careers. Universities Norway (UHR). https://www.uhr.no/en/_f/p3/i86e9ec84-3b3d-48ce-8167-bbae0f507ce8/nor-cam-a-tool-box-for-assessment-and-rewards.pdf

Widén, G., & Hansen, P. (2012). Managing collaborative information sharing: bridging research on information culture and collaborative information behaviour. Information Research, Vol. 17 No. 4, available at: https://informationr-1net-1unwr6w0y48e4.hps.bj.uj.edu.pl/ir/17-4/paper538.html (accessed September 14, 2024).

Yue, L., Ye, M., & Chen, Q. (2022). The impact of partnerships and information sharing on corporate sustainable performance: a mediation model moderated by government support. Frontiers in Psychology, 13, 942279. https://doi.org/10.3389/fpsyg.2022.942279

© CC-BY-NC 4.0 The Author(s). For more information, see our Open Access Policy.