ABOUT JDSR

JDSR is a interdisciplinary, online, open-access journal, focusing on the interaction between digital technologies and society. JDSR is published by DIGSUM, the Centre for Digital Social Research at Umeå University, Sweden.

JOURNAL OF DIGITAL SOCIAL RESEARCH

SILENCING TACTICS: PRONOUN CONTROVERSIES IN A COMMUNITY QUESTIONS AND ANSWERS SITE

Tanya K Osborne[a]

ABSTRACT

Understanding how and why online professional knowledge sharing communities develop issues with gender inclusion is essential to building safe and respectful environments. Trans and nonbinary gender identities are under constant threat and scrutiny, and trans people frequently face harms in online environments. Through digital ethnography, I explore how an international online programming community, Stack Exchange, responded to the challenges of implementing trans and nonbinary inclusive language policies. I discuss the rhetorical strategies and silencing tactics deployed by the community in response to policy changes. The analysis draws on Dotson's concept of testimonial smothering to argue that epistemic violence prevents dialogue about the importance of respecting preferred pronouns. The paper concludes with reflections on the implementation of pronoun policies in international communities.

Keywords: social epistemology, nonbinary gender, trans gender, epistemic violence, pronouns, stack exchange.

[a] Gothenburg University, Sweden.

1   INTRODUCTION

Stack Exchange is a network of community questions and answers websites, centred mainly around coding and programming. In late 2019, Stack Exchange introduced a new code of conduct which included an explicit statement about referring to others by their preferred pronouns. This change caused a sequence of events that still affects the community today. This paper follows the way that the community responded to these changes and explores the discourse about trans and nonbinary genders. I deploy the lens of epistemologies of ignorance in my analysis to demonstrate the effect of testimonial smothering in online environments. Testimonial smothering (Dotson, 2011) is a form of epistemic violence that causes people to silence their own experiences and accounts. In discussing these events, I uncover some of the concerns that affect international communities when negotiating pronoun policies and trans and nonbinary gender identities.

This research is guided by questions about how pronoun policies are implemented by online platforms, and what kinds of rhetorical strategies are adopted to discredit the validity of pronoun policies.

This paper builds on previous arguments about the relevance of epistemic injustice to misgendering (Argyriou, 2021), and situates this work in an empirical case. This paper also builds upon work done to understand the safety concerns of women in online knowledge sharing platforms (Menking et al., 2019), and extends this work to diverse gender identities. While other research has focussed on the extremes of geek masculinity (Regehr, 2020), in this paper I emphasise the ways in which the logics of geek masculinity and hegemonic masculinity are mainstreamed into professional knowledge sharing settings.

Throughout this work, I adopt the stance that trans and nonbinary identities have the potential to disrupt gender binaries (Butler, 2004), and adopt a critical stance toward gender binarism. I undertake ethnographic fieldwork with the experience of having worked in the software sector as a technical writer. In this context I was a regular reader of Stack Exchange. As someone who experiences gender as fluid, I am sensitised to the kinds of issues that arise in online spaces for people who identify as nonbinary.

Readers should be aware that the findings section of this paper contains reconstructions of hate speech towards trans and nonbinary people, and therefore may be triggering.

1.1 Stack Exchange

Established in 2009, Stack Exchange is a network of community questions and answers (QCA) forums covering a diverse range of topics. The flagship QCA forum on Stack Exchange is Stack Overflow, which is one of the largest online coding communities, with more than 15 million registered accounts and upwards of 50 million monthly users (Brooke, 2021; May et al., 2019). The company who run

Stack Exchange are called Stack Overflow, and they are based in the USA. Stack Overflow as a company not only operate the Stack Exchange network, but also run a careers site for coders and developers, and offer commercial knowledge sharing products. The focus of this paper is the Stack Exchange network and suite of QCA forums. This network cannot be treated as a single site with a single community; it is arguably an institutional setting that harbours many smaller communities and settings which are coordinated by Stack Exchange as an institution. An institution can be understood as a structure that uses texts to coordinate the activities of people (Smith, 1999, p. 196). Stack Exchange uses texts to coordinate the activity of users on the site, such as: blogs, the code of conduct, and platform mechanics such as reputation, and badges.

Knowledge sharing sites like Stack Exchange and Wikipedia, unlike other social media sites, grew from the commons-based and collaborative contentment creation movements (Menking et al., 2019; Reagle, 2013). The basic format of the site is a questions and answers forum, where users can upvote and downvote posts, and earn badges for their activities. Certain activities on the site generate reputation points, and accumulating reputation points opens up access to features of the site.

Stack Exchange dominates Google searches for information about coding and programming, and therefore, much like Wikipedia, has a strong epistemic power (Menking et al., 2019). Arguably, this site can be considered influenced by geek masculinities, both by its identity as a programming community, and by its situation as part of the creative commons movement (Menking et al., 2019).

Stack Exchange has a distinctive approach to moderation, and relies heavily on users moderating each other as peers. While the moderation approach does not easily map to the strategies of other large platforms, Stack Exchange uses a mixture of community-reliant volunteer moderation (Caplan, 2018) and automated moderation (Ponzanelli et al., 2014). Official Stack Exchange moderators are volunteers elected by their community who gain access to a moderator sub-community within the site. There are relatively few elected moderators on Stack Exchange compared to the active user base of the site: 24 moderators on Stack Overflow, and 540 across the entire Stack Exchange network (*Moderators - Stack Exchange*, 2021). However, any user can gain access to moderation tools by accumulating enough reputation points on the site. This means that most of the people who do moderation work are those who have earned enough reputation to access moderation tools, and not those who are elected as moderators. Those users do not necessarily have a strong connection to Stack Exchange as an institution. In other research, similar approaches to moderation have been linked with increases in alt-right opinions (Jasser et al., 2021).

Previous research on Stack Exchange has shown that the network has a longstanding issue with underrepresentation of women and under participation of women. This is evident on Stack Overflow (Nivala et al., 2020; Vasilescu et al., 2013), and it is also evident elsewhere on the Stack Exchange network. One study of the Graphic Design subsite, a field that typically has gender parity, estimated

that only 4% of the participants on the subsite were female (Dubois et al., 2020). Previous research specifically on female participation in Stack Exchange has relied on using computational models to determine the gender of posters, based on checking the username and profile picture for features that indicate gender (Brooke, 2021; Dubois et al., 2020; Ford, Harkins, et al., 2017; Vasilescu et al., 2013).

1.2 Trans and nonbinary experiences online

Social power regulates gender by reducing it to a binary of masculine or feminine identity, rendering other ways of relating to gender impossible or unthinkable (Butler, 2002). Gender identities outside of the binary offer opportunities to disrupt the social power that enforces the gender binary (Butler, 2004, p. 48). Trans and nonbinary identities are examples of such disruptive gender identities.

The internet offers a range of ways in which trans and nonbinary people are made visible, to each other and to the world. This can cause issues for trans people, who are more likely to experience harassment online compared to cis people (Powell et al., 2020). Harassment does not need to be targeted to cause harm; many trans people face "incidental harm" (Scheuerman et al., 2018) though witnessing content that is harmful but which is not directed at them personally.

Misgendering – or referring to someone in a way that does not respect their gender identity - is a common form of harassment that trans and nonbinary people experience online and offline. Misgendering is experienced chiefly through language and testimony, and can often be a result of linguistic conventions at a structural level (Argyriou, 2021). Being misgendered causes trans and nonbinary people to feel stigma, and previous research has suggested that this stigma is experienced most frequently by people who identify as genderqueer (McLemore, 2015). One way that misgendering can be prevented is through ensuring that people can state or display their pronouns. A study of trans and nonbinary college students found that people often appreciate the opportunity to declare their preferred pronouns in advance, so that they do not have to repeatedly come out (Goldberg et al., 2019). The same study also found that people who prefer the pronoun "they" experience the most resistance when asserting their preference of pronouns (Goldberg et al., 2019).

Being "out" in an online space also means dealing with the ways in which others impose stereotypes on people with LGBTQ+ identities (McKee, 2004). LGBTQ+ people may be strategic about the social networks where they are "out" about their sexuality or gender identity in order to avoid discrimination or harassment (Ford, Milewicz, et al., 2019; Talbot et al., 2020). However, the internet can also help trans and nonbinary people to explore their gender identities, for example, by allowing access to new concepts that describe their experiences (Scheuerman et al., 2018). For trans people, social media can provide a host of different ways to access information about the process of transitioning (Miller, 2017), and can be important for emotional support (Haimson, 2020).

1.3 Epistemic Violence and Testimony

In this paper, I explore how epistemic ignorance causes trans people to smother their testimony about their experiences online. Testimony can be understood as "any kind of telling in and through which the expression and transmission of knowledge becomes possible" (Medina, 2013, p. 28), which includes speech, writing, and digital communications that have the intent to convey the knowledge of the speaker.

Testimonial smothering is when a speaker withholds information from an exchange due to some external threat, and it is a form of silencing (Dotson, 2011). Dotson (2011) argues that testimonial smothering is a form of epistemic violence caused by pernicious ignorance. A pernicious ignorance arises from a reliable ignorance. Following Dotson's definitions, a "reliable ignorance" is an ignorance that arises from a known epistemic gap that is not necessarily harmful (Dotson, 2011). For example, it is possible to be ignorant about transgender people's experiences in the workplace, and be reflectively aware of that ignorance, without that ignorance causing harm. The ignorance might instead be a motivator to learn more. That ignorance becomes a "pernicious ignorance" when it can cause harm to someone in a given context. A pernicious ignorance might occur when an institution such as a large employer is ignorant to the harm of "deadnaming" (using a transgender person's pre-transition name) and that employer maintains a system that does not allow staff to update their names on their ID badges, forcing their transgender staff to be misgendered at work – a situation that occurred at Alphabet, Google's parent company (Mayo, 2021).

Dotson argues that a linguistic exchange leads to epistemic violence when an audience refuses to reciprocate in the exchange due to a pernicious ignorance. In certain circumstances, this kind of epistemic violence can lead to testimonial smothering. Dotson gives the following three linked circumstances as precondition to a speaker smothering their own testimony: (1) that the content of the testimony carries some risk to the speaker; (2) that the audience has demonstrated a testimonial incompetence to the speaker; and (3) that the testimonial incompetence arises from a pernicious ignorance (Dotson, 2011).

2 STUDY DESIGN

2.1 Methods

The data for this paper were drawn from ethnographic observation on the Stack Exchange network, herein referred to as SE. During ethnographic observation, key issues were identified from the problematic of gender, taking up the standpoint of nonbinary users and readers. Through observation, a research question was developed about the implementation of language inclusive policies and the rhetorical strategies used to discredit inclusive language. Observation was

conducted as an outsider and involved being present on the field site for four stretches during 2021. The observation focused initially on gender discourses on SE, and then developed a focus on the period during which the new code of conduct was implemented (Summer 2019 to Spring 2020). Having several shorter engagements allowed me to refine my observation strategies in response to my evolving understanding of the standpoint perspective. Analysis is based on memos written during observation and on the corpus of threads, documents, and other materials that were created through the observation. The issues were followed to other sites, social media, and blogs where they were being discussed, using a mobile approach to multi-sited ethnography (Hine, 2011), following the phenomenon and the discourse rather than individual subjects. Connections were traced through a wide variety of approaches, through exploration of the "online landscape" (Hine, 2007). In reconstructing the effects of implementing a language policy, a timeline approach was taken (Smith, 1990a). This involved comparing competing narratives about the introduction of the code of conduct and triangulating with other sources to compile an objective temporal ordering.

The main sites of observation were SE Meta, a subsite of SE where users discuss SE, blog posts written by SE, and blog posts written by SE users. However, the observation in total included a diverse cross section of the sub-sites and features of SE, and sites are discussed in the findings where relevant. Sites observed outside of SE included social media such as Reddit and Twitter; technology news sites; and technology blogs.

The strength of an ethnographic approach for this research is that it avoids some of the pitfalls in previous research on gender in SE. Previous research has relied on using usernames to predict gender, and researchers using these methods generally find that they cannot determine gender at all for between a third to a half of their samples (Ford, Harkins, et al., 2017; Vasilescu et al., 2013). Research on trans programmers has highlighted the inadequacy of name prediction approaches and the need to use approaches that consider the spectrum of gender (Ford, Milewicz, et al., 2019). By focusing on talk about gender rather than on the gender presentation of individual users, this research is able to develop a nuanced understanding about gender on SE.

2.2 Ethics

Where quotes are used in this paper, they are fictionalised composite quotes (Markham, 2012). These quotes may combine elements from several sources and are written to capture a typical exchange rather than one specific exchange. The reason for this is twofold; first to reduce searchability and thereby lessen the chance of identification; second to avoid focussing on individual interactions within the material, to keep the analysis at the social level. Usernames are also fictionalised and based on the general character of usernames on the site.

Material on SE is public and made available through a creative commons license, at time of writing this license is CC BY-SA v4 (*What Is the License for the Content I Post?*, 2021). No private material from SE, or from other websites and blogs, is quoted, and a sensitivity is taken toward which material might constitute "private" in these spaces (Markham & Buchanan, 2012). For example, while blog posts by SE users are publicly available, they are not quoted directly or indirectly in this paper.

3 FINDINGS

3.1 Telling Histories

In 2019, SE updated its code of conduct to specify that "be nice" means that users should refer to each other using their preferred pronouns, where they are stated. The events surrounding this change stirred up a huge controversy across the SE networks.

Unfolding the history of what transpired around these events reveals conflicts in the institutional sequencing of the timeline. Comparing between all these different retellings of the history of events, I reconstructed a timeline of my own to find an objective sequence of key events (Smith, 1990a).

Throughout my investigation of the history of the implementation of the new code of conduct, it seemed important to the community to recount the timeline of events, just as it seemed important for SE to offer some record of events themselves. There are a few key posts that are community written which contain a record of events and there is even a website dedicated to presenting a timeline of the events that transpired.

Most intriguing to me was that the majority of the community response to the new code of conduct happened before the new code of conduct was known and was based mostly on community hearsay about the content of the code of conduct. As the accounts of the history settle, the events are constructed to obfuscate the harms done to the trans community in favour of playing to a narrative about the institution losing touch with its core and acting against the interests of the community as a whole.

In my reconstruction of events, I start by positioning a few of the changes that SE were making to the business model of the site. In June 2019, paid advertisements were introduced to the site. Then, in September 2019, SE updated their creative commons licensing from CC BY-SA v3 to CC BY-SA v4. SE controversially applied this change in retrospect to contributions on the site. These are all changes that were negatively received by the community. Around this time, Stack Overflow appointed a new CEO. In user created timelines, these events were often positioned as evidence that the management of SE, and SE as an institution, are losing touch with its core userbase.

In September 2019, SE signalled to its moderators and staff that updates were incoming to the code of conduct. Immediately after this, one moderator resigned, citing in their resignation that they had been hurt by recent events, and that this was part of a larger pattern of harm towards trans and nonbinary people. The day after, a different moderator was fired outside of the established processes for the removal of moderators. The textual reconstruction of events frequently presented by other members of the community and by SE allow the reader to infer that the resignation and the firing are related. The firing was misattributed to bigoted behaviour, which is contested by the accounts of both involved moderators. This move is important to highlight and discuss; any solidarity with the trans community is quashed and replaced with vitriol that SE would fire a moderator unfairly. This is a subtle tactic for silencing a broader discussion about harms done to the trans community.

Following from this, other moderators started to resign in sympathy. The first sympathy resignations began in late September 2019 and continue until time of writing. From an analysis of the 35 resignations that took place between 28th September 2019 and 31 December 2019, only two mentioned the violence toward the LGBTQ+ community, 24 mentioned the moderator who was fired, and 8 specifically mentioned that they disagreed with the new code of conduct. Of those 8, three are very overt in saying that they disagree with referring to people by their preferred pronouns. These kinds of statements fed into a narrative that later emerges outside of SE, which positioned this incident as a case of political correctness gone awry, or a case of forcing Christian groups to accept LGBTQ+ framings that curtail their freedom of religious expression. Other resignations stated that their reasons for resigning were the changes to site licensing, or general concerns with the quality of the site. These statements fed into a contesting narrative that emerged within SE about the incoming management changes exploiting the userbase for financial reasons. A community written timeline springs up, which was frequently cited in these resignation posts. That timeline omitted the story of the original resigned moderator from the retelling and did not mention anything about the harms done to the trans community. It seems that the community found it easier to side-line the potentially difficult conversation about the treatment of trans and nonbinary people within SE in favour of rallying around the common cause of the way in which one moderator was unfairly fired.

At or around this time, users started to change their usernames in support of the fired moderator, adding phrases like "Reinstate the moderator" to their names. Eleven days after the original incident, SE made an official response. At this point, the issue at hand was known to be related to the upcoming code of conduct changes, and was presumed to be about pronouns, but the code of conduct was not yet public. Members of the community created a "Pronoun Assist" script in October 2019, described in more detail later in this paper.

Members of the moderator community wrote two letters to SE on SE Meta in October 2019, titled *Dear Stack Exchange,* which focussed on the harms done

to SE's relationship the community, and *The Lavender Letter*, which focussed on the harms done to the LGBTQ+ community on SE, specifically on homophobic and transphobic incidents. These letters were open for anyone to sign to show their support.

The new code of conduct was officially published on 10 October 2019, accompanied by an FAQ. The original FAQ received a large number of downvotes and attracted a lot of vicious transmisogynistic comments, and it was closed down and replaced with a new version on 22 October 2019.

The fired moderator eventually successfully issued a legal challenge to SE. Around November 2019 SE announced that community ads can no longer be used to promote legal fundraisers, which implies that there was some large-scale activity at this time to promote the crowdfund campaign. As a researcher, these kinds of actions, combined with the number of users who had changed their names in solidarity, lead me to investigate what had happened.

In the months after, SE made the first tentative steps towards repairing the damage done to the community, through backtracking on changes to creative commons licensing (March 2020), introducing diversity training for elected moderators (June 2020) and responding to *The Lavender Letter* (October 2020).

3.2 Reflections on Histories

In the official textual construction of this event, there is a complete "textual silence" (Huckin, 2002) about the very fact that the issue at hand (pronouns) is raised in relation to trans and nonbinary identities. In the code of conduct, the imperative to use stated pronouns is broadly outlined under the heading of "no bigotry". The relevance of trans and nonbinary identities is instead made clear in the FAQs. FAQs on SE are presented in a similar manner to their main questions and answer content. SE distances themselves from the responsibility of contextualising the reasoning behind referencing preferred pronouns, and places that responsibility in the hands of its community. In not explaining the issue, SE avoid themselves taking a stance or making their stance explicit. This is something that happens frequently when companies perform lip service to LGBTQ+ inclusivity, which is a phenomenon sometimes called "rainbowwashing" (Wolowic et al., 2017; Wulf et al., 2022).

When we follow the debate away from the platform and towards technology news sources, the positioning of trans and nonbinary identities becomes clear. The firing of a moderator is called "punishment for crimes against 'wokeness'" in the more alt-right leaning news comments. Invocation of "woke" alludes sarcastically to what is seen as a left-wing youth movement that has a performative approach to inclusion, implying that performing "wokeness" is about appearing morally superior without doing any of the work. If the traditionally right wing, anti-trans corners of the internet think that these actions are produced as part of a theatre of inclusion, it may be a surprise that LGBTQ+ communities within SE agree in some sense

with this position. In these circles, while the rhetoric of "woke" isn't invoked, the conversation is about how the policy conceived by SE is only a lip service the real work and doesn't go far enough to meet the needs of the community.

From this, two different discursive moves become apparent. The first discursive move occurs on SE. In this move, the story of the original resignation is gradually erased from the various accounts, and instead the change to the code of conduct is framed as the disruptive force that symbolises the growing disconnect between SE as a business and SE as a community. The second discursive move happens outside of SE. This move lionizes the fired moderator and positions them as an oppressed voice of reason and a lone warrior against so-called politically correct language movements – perhaps against their own intentions.

3.2.1 *Global Contexts*

Iterations of the discussion about the code of conduct on SE Meta make clear that one thing was sorely lost in the production of the code of conduct: how those with English as a second or other language understand pronouns and the politics of pronouns. In the below example, I present a typical exchange:

> Tomer: In my native language, the same pronoun is used for both he and for gender neutral purposes. I sometimes have to correct myself, but I still try.
>
> AndrewL: Imagine how hard this is for speakers of languages that don't even have pronouns! At least you know where to start. SE can't teach every user how to write in English.
>
> AI_dragon: Sometimes I see posts where the grammar is so bad, I don't think users are capable of understanding being corrected on their pronouns. We'd have to ban everyone.
>
> null: Average users will be so confused by all this

Conversation re-enacted from SE Meta

Those who speak English as a second language were learning for the first time that "he" is not a gender-neutral term or reveal that they were taught to use "he" as a gender-neutral term as part of their English language education. To a specifically US and anglophone audience, the subtext of introducing this pronoun discourse was that of inclusivity, but without a knowledge of the US context this subtext becomes difficult to grasp. It represents a reliable ignorance (Dotson, 2011), but not a pernicious ignorance, on behalf of users who are speaking English as a second language. It is a reliable ignorance because we should expect that an audience outside of an anglophone context does not know about the contemporary issues around pronouns.

The audience for SE is global, with results from a survey of users suggesting that the top three nations who use the site are: the US, with approximately 20% of users; followed by 13% from India, and 6% from Germany (*Stack Overflow Developer Survey*, 2020). Despite this, the discourse of the site makes apparent the assumption that everyone shares a US frame of reference for societal issues. This ignoring of the global context allows a hermeneutical injustice (Medina, 2013, p. 91) to surface, as the parties involved in conversation lack access to the same conceptual knowledge.

Global contexts include a wide variety of different understandings and approaches to gender and gender roles. For example, some cultures and contexts already include concepts for diverse and nonbinary genders, the existence of which are often erased or changed through colonialism (Benson, 2020; Chatterjee, 2018). The second largest audience for SE is India, as noted above. India and South Asia have a unique cultural context for gender diversity, and recognise a variety of gender identities outside of the westernised binary, with varying degrees of legal protection (Dutta et al., 2019). The use of the word trans in the Indian discourse is contested, carrying substantial colonialist baggage (Chatterjee, 2018), with the preference being towards using gender diverse as the umbrella term (Jain & Rhoten, 2020). Despite the strong presence of India in SE, there didn't seem to be room for Indian and South Asian perspectives to emerge within the discourse around pronoun policies.

3.2.2 The Risk of Speaking Up

A clear case of testimonial smothering emerges from this debate. *The Lavender Letter* states that the nature of the debate around pronouns caused harms to the trans and nonbinary community, at times in the form of direct bullying and harassment. However, the risk for someone within that community to speak outwardly on SE Meta about their experiences is high.

In this environment, risks can have very real consequences, for example, posting something that is heavily downvoted can cause reputation loss and could therefore cause the poster to lose access to important site functions. For those who use their real identities when posting, as is encouraged by the entanglement of SE with social media and job markets, this can also pose genuine threat to personal security. Being identifiable in real life from such posts could easily lead to trans people encountering physical threats and harassment in the real world (Scheuerman et al., 2018).

As previous interactions about the matter are very visible, a user can quickly assess for themselves how likely it is that the audience is competent to hear their testimony. If they see mostly repetitions of familiar bigoted strategies, they will surmise that the audience is incompetent. In this case the audience is incompetent because it has a reliable ignorance about the importance of pronouns within the trans and nonbinary communities. This ignorance is pernicious because it can cause

harm to individuals by resulting in misgendering. When a user is aware that the message they want to send will be rejected by the community, they are likely to simply not post and therefore are subject to smothering their own testimony (Dotson, 2011).

3.2.3   User interventions

SE contains and encourages multiple ways for users to interact with the API and data that is housed on the platform. Among these is a subsite known as Stack Apps, where users can share scripts and browser extensions that they have written in order to in some way extend or modify their experiences of the site. These scripts can be understood as ways in which the users resist the affordances of the site and co-create their experiences and opportunities on the site by applying their technical expertise.

Popular user scripts are varied and creative, for example, one automates writing common comments, one allows you to virtually "punch" annoying users. Among the most popular of these scripts is an extension that allows users to display the preferred pronouns of other users next to their usernames on posts.

The post that presents the instructions for installing this script directly references the incoming code of conduct, so we could see this as a community response. This oppositional action shows that what the community want is to be able to exercise some active agency over the terms by which their pronouns and known and used. If it were a standard function of the site to enable users to prominently display their pronouns, it would also enable the community to hold others accountable for ignoring this information. My experience of using this script while browsing the site was that while it was relatively rare to spot someone stating preferred pronouns on Stack Overflow, it was more common on the non-technology oriented subsites of SE, such as SE Meta.

3.3   Common anti-trans argumentative strategies

In this part of the analysis, I put forward two of the more common anti-trans argumentative strategies that users invoke: constructed imperilment and compelled speech. These argumentative strategies are in no way unique to anti-trans rhetoric, but will be familiar to those studying hate speech against many marginalised groups. The similarities between these tactics when used in transphobic settings and in racist, anti-immigrant, and nationalist settings are indicative of power relations that benefit from constructing some lives as more worthy than others (Snorton, 2013).

3.3.1   Constructed Imperilment

Something I commonly observed in discussions was an appeal to ways in which people might be wronged by using the correct pronouns. The effect of these argumentative moves is to imply a hierarchy of being wronged and prevent others

from asserting their preferences. It purposefully ignores that people might occupy multiple marginalised identities. The following example shows a typical exchange:

> RegEx_Fan: Some people might be made really uncomfortable by having to use certain pronouns, and those people could have that problem on account of having autism or being a good Christian. Why do we prioritise the discomfort of trans people?
>
> HappyShark If I don't feel comfortable saying "they", I should be allowed to not write pronouns at all.
>
> Fimbrethil: Totally agree.
>
> RegEx_Fan: It seems like making a mountain out of a molehill, not like we can tell who anyone is online anyway.

Conversation re-enacted from SE Meta

The typical focus for discomfort in these conversations is on the use of the pronoun "they" as a singular personal pronoun, and it was very rare to see other pronoun strategies mentioned. In these exchanges, using "they" in this way is imagined to cause difficulties for either religious or personal reasons.

The argumentative moves shown above fit well with the "constructed imperilment" (Marcks & Pawelz, 2020) tactics also seen in far-right discourse; it is a movement towards denying trans people's rights to exist by implying that their existence endangers the existence of others. Constructed imperilment has been studied in the context of anti-immigration rhetoric (Marcks & Pawelz, 2020), but here we see the same kinds of rhetorical strategies used against pronouns.

If we accept that others might be wronged by using the correct pronouns, we must then choose who we wrong, and who it is more important to not wrong. This falsely presents a situation wherein asking to be respected constitutes a wrong. This is an argumentative move toward pre-emptively silencing someone from making their preferences known. It is also possible to observe in these interactions a fear of the disruptive other. Introducing these disruptive others disrupts the comfort of male hegemony by presenting a reminder of contested and divisive real-world issues in otherwise sheltered online bubbles (Nakamura, 2002, p. 37).

3.3.2 The "Compelled Speech" Argument

A commonly repeated argument against mandating the use of preferred pronouns is that would constitute a class of "compelled speech". An appeal to "compelled speech" is specifically an appeal to the US 1st Amendment, which protects the freedom of speech for US citizens. The below reconstruction is typical of an exchange on the matter. Very similar exchanges occurred on SE Meta, and in other sources.

DiverChuck: I don't want to be in a community where I have to use modern pronouns. It's against my preferences. No way I'm making an account there.

Chris72: It should be ILLEGAL. It is COMPELLED SPEECH. They FORCE people to LIE.

STEMhamster: The "activists" should be helping to protect our freedom of speech instead of wasting our time by inventing a new gender every day.

Conversation re-enacted from comments on technology news websites

In recent lawsuits, the US 1st Amendment has been used to argue for the right to not use trans inclusive language, and has not succeeded (Eckes, 2021). Appealing to laws such as this may offer an individual protection to speak as they wish, but they do not offer protection from other laws that are targeted to prevent discrimination.

A question remains as to whether implementing a pronoun policy really can be understood as compelled speech. From a social epistemological point of view, we can view compelled speech as speech that has been elicited from an agent, where the speech does unjust harm against the agent (McKinney, 2016). The flaw to arguing that a pronoun policy elicits unjust speech in the context of SE is twofold. Firstly, such speech is not coerced or manipulated from the poster. There is always a possibility of writing an exchange in a professional format that does not make use of pronouns at all. Secondly, all questions and answers on SE are to some extent co-owned and editable by anyone else, making literal forms of compelled speech – someone changing your own words – inevitable. Regardless, a poster always retains the ability to delete a post or remove their username from a post if it is too popular to be deleted. There is always a way to retain control over the speech act if the poster feels that they are harmed by the way that their speech has been altered. In this case, the speech does not seem to meet the requirements to be compelled speech.

Recourse to "freedom of speech" has been noted elsewhere as a common strategy to avoid using preferred pronouns by those that identify as within the alt-right and men's rights movements (Haslop & O'Rourke, 2021). Significant in the case of SE is that the same discourse is used outside of these contexts. Whether this indicates a deeper relation of ruling between the extreme men's rights movements and geek masculinities is uncertain. Similar relationships have been noted between anti-trans hate campaigns and liberal democratic positions (Gill-Peterson, 2021). Arguably, such alt-right and anti-trans frames act as a coordinating force in these discussions, with the exact same arguments and logics repeating themselves in professional contexts.

Situations like this open up to "mixed legibility", a common feature of microaggressions (Schroer & Bain, 2020). Utterances with mixed legibility are instances where statements are deliberately intended to convey different messages to different audiences, in a way that enables the speaker to plausibly deny any intended harm (Schroer & Bain, 2020). Some may choose to read these appeals to

freedom of speech as a fair-minded, classically liberal approach to public discourse; but those trans and nonbinary people who are familiar with this rhetoric and their harms will understand it as an attack on their position as knowers. These attacks on the embodied knowledge of gender held by trans and nonbinary people presents misgendering as simply exercising free speech.

3.4 The Burden of Proof

Ignorance about women and gender minorities experiences online and in the workplace is a reliable ignorance that can become a pernicious ignorance when it is used to stifle debates that are invoked to improve those experiences. In many contexts, the primary mechanism by which such debates are shut down is a recourse to the burden of proof.

"Proof" is itself a discourse that surfaces on SE about discussions on discrimination or exclusion. Vocal, usually male-presenting users demand "proof" that women and gender minorities have a worse experience on the site. Individual experience, within this discourse, is not considered "proof". It is not always clear what would constitute satisfactory "proof", but the consensus appears to be that "data" would be considered proof. Commentors suggest that admissible "data" might be text mined from posts in a uniform, "objective" way. However, "data" from the internal Stack Overflow survey that shows women, gender minorities and people of colour saying in larger numbers that they do not feel that they belong to the Stack Overflow community (*Stack Overflow Developer Survey*, 2020) is not considered admissible "data" in support of this claim.

I suggest that this kind of argument shows "data" as part of an ideological ordering of the social (Smith, 1990b, p. 145). From this perspective, in order for something to go from observation to fact it must be rendered objective through "data" (Smith, 1990b, p. 45), the possibility of knowledge that arises from situated embodiment is completely closed off.

Not only do woman and gender minorities need to be resilient to experiences of harmful behaviours, but they also have to "prove" that these behaviours happen if they want to speak about them. Limiting the concept of "proof" to a very rigid and perhaps unattainable set of standards is a way of silencing talk about what discrimination means in this environment. By some accounts of epistemic ignorance, this kind of action could be considered a way of discrediting testimony before receiving it (Mills, 2017), which can also be understood as a form of testimonial injustice.

In addition to that, people who are harmed by these behaviours are also expected to be the ones who correct it. Reading posts by members of the LGBTQ+ community shows repeatedly that people feel that the institution stirs up controversies, then abandons potentially vulnerable people to speak out and police behaviour themselves. We see this as evidenced by the timeline, with SE outsourcing the FAQs for the code of conduct to the community, expecting that

trans people and their allies will fill in the silence about the meaning of pronouns. This is a problematic positioning, and does not consider the safety of the community (Menking et al., 2019).

There have been numerous incidents on SE of male posters creating female-presenting alternative accounts in order to gather evidence about the state of discrimination on the site. While some have chosen to remain silent about their experience, there is one well known instance where the user posted a detailed breakdown of their interactions on the site as a female user, concluding that his experience was the same, if not better, than his experience as a male-presenting user. The user claimed that in many cases the community was more willing to help them under their new female guise. This post was presented as "proof" that the site has no issues being welcoming to women.

While these behaviours might be understood as "identity tourism" (Nakamura, 2002, p. 40), the recreational appropriation of identity in order to take a vacation from one's real life, I would argue that this is not the case here. These incidents take the shape of espionage, a secret mission to capture the view of the other and expose their perceived lies. They are not a vacation; they are a fact-finding mission.

For women and trans people wanting to challenge this idea of proof, the task may seem impossible. The risks of engaging in this debate are twofold. One - for some, engaging this debate means coming out as a female, trans, or gender diverse user on the site, who might be choosing to occupy a gender neutral, or even deliberately masculine user persona. Two – with this disclosed, the user potentially faces harassment from others for occupying a marginalised gender position, as is clearly demonstrated when others have tried to explain what the gender diverse user experience is like. We see here a pernicious ignorance among the user base. They are reliably ignorant about the experiences of others but are also excessively defensive about the possibility of a difference in experience. Their need to preserve their ignorance about misogyny and harassment faced by gender diverse users in these spaces causes them to violently shut down these conversations.

4   LIMITATIONS

This paper concentrates on the implementation of a code of conduct in one professional knowledge sharing setting, which is heavily weighted toward male participation. Communities that are less international and that have a different balance of gender might encounter completely different challenges in implementing pronoun policies.

Further research in this area could be enriched by including interview data. In the context of this paper, I felt that using interview data was ethically problematic, as it may have presented issues of privacy violation to participants who occupy vulnerable positions (Korn, 2019).

5   CONCLUSION

In this paper, I have explored the implementation of a pronoun policy, and examined the rhetoric and tactics used to discredit pronoun policies. I found that the rhetorical strategies that are used to discredit pronoun policies draw on the same argumentative concepts often seen in alt-right settings, and have marked similarities to the strategies of hate that are used against other marginalised groups.

In the case of SE, a conversation that should have been about an online platform can do justice for trans people became obscured by a debate about how an institution should balance priorities between its business model and its volunteer moderator staff. In this way, pernicious ignorance is mobilised an institutional setting as a tactic to silence oppositional discourse.

Strikingly, this paper shows trans and nonbinary people cast in the role of the disruptive other, bringing with them the threat of expulsion for not following the pronoun rules.

Among the various silencing tactics, this paper reflected on ways in which different ideological deployments of the notions of proof can distort opportunities for discussion by purposefully devaluing embodied experiences. Such renderings of data as the ultimate bearer of truth are ultimately used to silence and stifle discussion about lived experiences. These discourses are often complicit in preventing marginalised groups from making epistemological contributions.

Transmisogyny has long been recognised as a phenomenon, but intersectional understandings of anti-trans hate speech often fall by the wayside (Scheuerman et al., 2018). In this paper, I find that hate speech is directed equally to trans people and to people who prefer to use the gender neutral pronoun "they". In line with Scheuerman et al (2018), this study finds that there are missed opportunities to design and co-create policies that disrupt intersectional marginalisation. An inadequate understanding of how gender and cultural dynamics play out on SE led ultimately to an implementation of policy that caused harm, which might have been avoided by properly engaging with the community.

Other research has found that low levels of content moderation and pseudo-anonymity are key ingredients in helping alt-right spaces to flourish (Jasser et al., 2021), and has shown that alt-right discourse uses emotions to mobilise hateful actions (Marcks & Pawelz, 2020). In this paper, I demonstrate that these tactics exist in internet settings that are not considered radicalised or aligned with the alt-right. While this paper could not investigate the relationship of coordination between sites like Stack Exchange and sites like 4Chan or Reddit, future work could focus on the bleed through and propagation of alt-right views into the everyday internet.

This paper raises important questions about pronoun policies that require further research and investigation. Firstly, a question arises regarding what kind of information should be included in a pronoun policy. For example, within the data corpus of this paper, I found that neopronouns (such as ze or ey) were seldom

mentioned, and other pronoun strategies used by queer people (such as dropping pronouns entirely) were mostly absent from discussions. Further research could explore introducing other pronoun strategies to online communities. Secondly, a question arises regarding intercultural communication and pronouns. There exists a strong tension between the dominant westernised, anglophone interpretations of trans and nonbinary gender issues, and the manyfold, nuanced expressions of trans and nonbinary gender issues on the global stage. These tensions deserve research attention in order to improve approaches to inclusivity in large, international online communities.

FUNDING STATEMENT AND ACKNOWLEDGMENTS

REFERENCES

Argyriou, K. (2021). Misgendering as epistemic injustice. Las Torres de Lucca. International Journal of Political Philosophy, 10(19), 71–82. https://doi.org/10.5209/ltdl.76464

Benson, K. (2020). What's in a Pronoun?: The Ungovernability and Misgendering of Trans Native Kids in Juvenile Justice in Washington State. Journal of Homosexuality, 67(12), 1691–1712. https://doi.org/10.1080/00918369.2019.1613854

Brooke, S. J. (2021). Trouble in programmer's paradise: Gender-biases in sharing and recognising technical knowledge on Stack Overflow. Information, Communication & Society, 0(0), 1–22. https://doi.org/10.1080/1369118X.2021.1962943

Butler, J. (2002). Gender Trouble: Tenth Anniversary Edition. Routledge. https://doi.org/10.4324/9780203902752

Butler, J. (2004). Undoing Gender. Taylor & Francis Group. https://doi.org/10.4324/9780203499627

Caplan, R. (2018). Content or Context Moderation? Data & Society. https://datasociety.net/library/content-or-context-moderation/

Chatterjee, S. (2018). Transgender Shifts: Notes on Resignification of Gender and Sexuality in India. TSQ: Transgender Studies Quarterly, 5(3), 311–320. https://doi.org/10.1215/23289252-6900696

Dotson, K. (2011). Tracking Epistemic Violence, Tracking Practices of Silencing. Hypatia, 26(2), 236–257. https://doi.org/10.1111/j.1527-2001.2011.01177.x

Dubois, P. M. J., Maftouni, M., Chilana, P. K., McGrenere, J., & Bunt, A. (2020). Gender Differences in Graphic Design Q&As: How Community and Site Characteristics Contribute to Gender Gaps in Answering Questions. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 113:1-113:26. https://doi.org/10.1145/3415184

Dutta, S., Khan, S., & Lorway, R. (2019). Following the divine: An ethnographic study of structural violence among transgender jogappas in South India. Culture, Health & Sexuality, 21(11), 1240–1256. https://doi.org/10.1080/13691058.2018.1555718

Eckes, S. E. (2021). Pronouns and Preferred Names: When Public School Teachers' Religious Beliefs Conflict With School Directives. Educational Researcher, 50(1), 65–68. https://doi.org/10.3102/0013189X20943198

Ford, D., Harkins, A., & Parnin, C. (2017, November 20). Someone Like Me: How Does Peer Parity Influence Participation of Women on Stack Overflow? 2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). https://doi.org/10.1109/VLHCC.2017.8103473

Ford, D., Milewicz, R., & Serebrenik, A. (2019). How Remote Work Can Foster a More Inclusive Environment for Transgender Developers. 2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE), 9–12. https://doi.org/10.1109/GE.2019.00011

Gill-Peterson, J. (2021, September 13). From Gender Critical to QAnon: Anti-Trans Politics and the Laundering of Conspiracy. The New Inquiry. https://thenewinquiry.com/from-gender-critical-to-qanon-anti-trans-politics-and-the-laundering-of-conspiracy/

Goldberg, A. E., Beemyn, G., & Smith, J. Z. (2019). What Is Needed, What Is Valued: Trans Students' Perspectives on Trans-Inclusive Policies and Practices in Higher Education. Journal of Educational and Psychological Consultation, 29(1), 27–67. https://doi.org/10.1080/10474412.2018.1480376

Haimson, O. L. (2020). Challenging "Getting Better" Social Media Narratives with Intersectional Transgender Lived Experiences. Social Media + Society, 6(1), 2056305120905365. https://doi.org/10.1177/2056305120905365

Haslop, C., & O'Rourke, F. (2021). 'I mean, in my opinion, I have it the worst, because I am white. I am male. I am heterosexual': Questioning the inclusivity of reconfigured hegemonic masculinities in a UK student online

culture. Information, Communication & Society, 24(8), 1108–1122. https://doi.org/10.1080/1369118X.2020.1792531

Hine, C. (2007). Connective Ethnography for the Exploration of e-Science. Journal of Computer-Mediated Communication, 12(2), 618–634. https://doi.org/10.1111/j.1083-6101.2007.00341.x

Hine, C. (2011). Towards ethnography of television on the internet: A mobile strategy for exploring mundane interpretive activities. Media, Culture & Society, 33(4), 567–582. https://doi.org/10.1177/0163443711401940

Huckin, T. (2002). Textual silence and the discourse of homelessness. Discourse & Society, 13(3), 347–372. https://doi.org/10.1177/0957926502013003054

Jain, D., & Rhoten, K. M. (2020). Epistemic Injustice and Judicial Discourse on Transgender Rights in India: Uncovering Temporal Pluralism. Journal of Human Values, 26(1), 30–49. https://doi.org/10.1177/0971685819890186

Jasser, G., McSwiney, J., Pertwee, E., & Zannettou, S. (2021). 'Welcome to #GabFam': Far-right virtual community on Gab. New Media & Society, 14614448211024546. https://doi.org/10.1177/14614448211024546

Korn, J. U. (2019). #IfTheyGunnedMeDown: How Ethics, Gender, and Race Intersect When Researching Race and Racism on Tumblr. Journal of Digital Social Research, 1(1), 41–44. https://doi.org/10.33621/jdsr.v1i1.12

Marcks, H., & Pawelz, J. (2020). From Myths of Victimhood to Fantasies of Violence: How Far-Right Narratives of Imperilment Work. Terrorism and Political Violence, 0(0), 1–18. https://doi.org/10.1080/09546553.2020.1788544

Markham, A. (2012). Fabrication as Ethical Practice. Information, Communication & Society, 15(3), 334–353. https://doi.org/10.1080/1369118X.2011.641993

Markham, A., & Buchanan, E. (2012). Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0) (p. 19). Association of Internet Researchers. http://www.aoir.org/reports/ethics2.pdf

May, A., Wachs, J., & Hannák, A. (2019). Gender differences in participation and reward on Stack Overflow. Empirical Software Engineering, 24(4), 1997–2019. https://doi.org/10.1007/s10664-019-09685-x

Mayo, A. (2021, June 2). Google workers are pushing the company to use preferred names on ID badges and drop deadnames. Business Insider. https://www.businessinsider.com/alphabet-union-google-workers-id-name-change-deadnames-trans-employees-2021-6

McKee, H. (2004). "Always a shadow of hope": Heteronormative binaries in an online discussion of sexuality and sexual orientation. Computers and Composition, 21(3), 315–340. https://doi.org/10.1016/j.compcom.2004.05.002

McKinney, R. A. (2016). Extracted Speech. Social Theory and Practice, 42(2), 258–284. https://doi.org/10.5840/soctheorpract201642215

McLemore, K. A. (2015). Experiences with Misgendering: Identity Misclassification of Transgender Spectrum Individuals. Self and Identity, 14(1), 51–74. https://doi.org/10.1080/15298868.2014.950691

Medina, J. (2013). The epistemology of resistance gender and racial oppression, epistemic injustice, and resistant imaginations. New York : Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199929023.001.0001

Menking, A., Erickson, I., & Pratt, W. (2019). People Who Can Take It: How Women Wikipedians Negotiate and Navigate Safety. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300702

Miller, B. (2017). YouTube as Educator: A Content Analysis of Issues, Themes, and the Educational Value of Transgender-Created Online Videos. Social Media + Society, 3(2), 2056305117716271. https://doi.org/10.1177/2056305117716271

Mills, C. W. (2017). White Ignorance. In Black Rights/White Wrongs (pp. 49–71). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190245412.003.0004

Moderators—Stack Exchange. (2021, June 29). Stack Exchange. https://stackexchange.com/about/moderators

Nakamura, L. (2002). Cybertypes: Race, ethnicity, and identity on the Internet. Routledge. https://doi.org/10.4324/9780203699188

Nivala, M., Seredko, A., Osborne, T., & Hillman, T. (2020). Stack Overflow – Informal learning and the global expansion of professional development and opportunities in programming? 2020 IEEE Global Engineering Education Conference (EDUCON), 402–408. https://doi.org/10.1109/EDUCON45650.2020.9125165

Ponzanelli, L., Mocci, A., Bacchelli, A., Lanza, M., & Fullerton, D. (2014). Improving Low Quality Stack Overflow Post Detection. 2014 IEEE International Conference on Software Maintenance and Evolution, 541–544. https://doi.org/10.1109/ICSME.2014.90

Powell, A., Scott, A. J., & Henry, N. (2020). Digital harassment and abuse: Experiences of sexuality and gender minority adults. European Journal of Criminology, 17(2), 199–223. https://doi.org/10.1177/1477370818788006

Reagle, J. (2013). "Free as in sexist?" Free culture and the gender gap. First Monday. https://doi.org/10.5210/fm.v18i1.4291

Regehr, K. (2020). In(cel)doctrination: How technologically facilitated misogyny moves violence off screens and on to streets. New Media & Society, 1461444820959019. https://doi.org/10.1177/1461444820959019

Scheuerman, M. K., Branham, S. M., & Hamidi, F. (2018). Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm

with Transgender People. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 155:1-155:27. https://doi.org/10.1145/3274424

Schroer, J. W., & Bain, Z. (2020). The Message in the Microaggression: Epistemic Oppression at the Intersection of Disability and Race. In Microaggressions and Philosophy (p. 25). Routledge. https://doi.org/10.4324/9780429022470

Smith, D. E. (1990a). Texts, facts, and femininity: Exploring the relations of ruling. Routledge. https://doi.org/10.4324/9780203425022

Smith, D. E. (1990b). The conceptual practices of power: A feminist sociology of knowledge.

Smith, D. E. (1999). Writing the social: Critique, theory, and investigations. Univ. of Toronto Press. https://doi.org/10.3138/9781442683747

Snorton, C. R. (2013). Marriage Mimesis. Journal of International and Intercultural Communication, 6(2), 127–134. https://doi.org/10.1080/17513057.2013.776095

Stack Overflow Developer Survey. (2020). Stack Overflow. https://insights.stackoverflow.com/survey/2020/?utm_source=social-share&utm_medium=social&utm_campaign=dev-survey-2020

Talbot, C. V., Talbot, A., Roe, D. J., & Briggs, P. (2020). The management of LGBTQ+ identities on social media: A student perspective. New Media & Society, 1461444820981009. https://doi.org/10.1177/1461444820981009

Vasilescu, B., Capiluppi, A., & Serebrenik, A. (2013). Gender, Representation and Online Participation: A Quantitative Study. Interacting with Computers, 26. https://doi.org/10.1093/iwc/iwt047

What is the license for the content I post? - Help Center. (2021). Meta Stack Exchange. https://meta.stackexchange.com/help/licensing

Wolowic, J. M., Heston, L. V., Saewyc, E. M., Porta, C., & Eisenberg, M. E. (2017). Chasing the rainbow: Lesbian, gay, bisexual, transgender and queer youth and pride semiotics. Culture, Health & Sexuality, 19(5), 557–571. https://doi.org/10.1080/13691058.2016.1251613

Wulf, T., Naderer, B., Olbermann, Z., & Hohner, J. (2022). Finding gold at the end of the rainbowflag? Claim vagueness and presence of emotional imagery as factors to perceive rainbowwashing. International Journal of Advertising, 0(0), 1–21. https://doi.org/10.1080/02650487.2022.2053393

COPING WITH ALGORITHMIC RISKS: HOW INTERNET USERS IMPLEMENT SELF-HELP STRATEGIES TO REDUCE RISKS RELATED TO ALGORITHMIC SELECTION

Kiran Kappeler[a], Noemi Festic[a], Michael Latzer[a] and Tanja Rüedy[a]

ABSTRACT

Algorithmic selection is omnipresent in various domains of our online everyday lives: it ranks our search results, curates our social media news feeds, or recommends videos to watch and music to listen to. This widespread application of algorithmic selection on the internet can be associated with risks like feeling surveilled (S), feeling exposed to distorted information (D), or feeling like one is using the internet too excessively (O). One way in which internet users can cope with such algorithmic risks is by applying self-help strategies such as adjusting their privacy settings (S_{strat}), double-checking information (D_{strat}), or deliberately ignoring automated recommendations (O_{strat}). This article determines the association of the theoretically derived factors risk awareness (1), personal risk affectedness (2), and algorithm skills (3) with these self-help strategies. The findings from structural equation modelling on survey data representative for the Swiss online population (N_{2018}=1,202) show that personal affectedness by algorithmic risks, awareness of algorithmic risks and algorithm skills are associated with the use of self-help strategies. These results indicate that besides implementing statutory regulation, policy makers have the option to encourage internet users' self-help by increasing their awareness of algorithmic risks, clarifying how such risks affect them personally, and promoting their algorithm skills.

Keywords: algorithmic risks; coping; self-help strategies; governance of algorithms; governance choice; survey method

---

[a] University of Zurich, Switzerland.

1   INTRODUCTION

An increasing share of our daily lives is spent online with endless options offered for a broad range of our needs, such as entertainment, information seeking, or socializing. To catch and maintain internet users' attention, online services (e.g., search engines, social media, streaming platforms) draw on algorithms to select the content with the highest probability of matching their users' interests. This algorithmic selection increasingly governs our daily lives, for instance, by ranking search results or news articles, recommending a certain movie, or curating one's social media feed (Festic, 2020; Latzer et al., 2016; Latzer & Festic, 2019). Hence, algorithms embedded in widespread online services affect our daily lives in many ways by automatically selecting pieces of information and assigning relevance to them (Latzer & Just, 2020). An input-throughput-output model helps to better grasp this understanding of algorithmic selection (Latzer et al., 2016): based on *input* data (e.g., users' click behavior, user requests), computational procedures (*throughput*) produce an algorithmically selected *output* (e.g., tailored news feeds, personalized recommendations).

While the widespread application of algorithms on the internet brings benefits like the reduction of complexity, it also entails challenges and risks for individuals and society as a whole (Just & Latzer, 2017; Latzer et al., 2016). Algorithmic selection relies on user data that is constantly being collected (Büchi et al., 2020; Hildebrandt, 2008). In many instances, internet users are not consciously sharing their data and are thus not aware of the data traces they produce (Micheli et al., 2018). The algorithmic processing of this collected data entails risks.

So far, research on algorithmic risks and how internet users cope with them predominantly focused on privacy (e.g., Boerman et al., 2018; Büchi et al., 2017). As a consequence of the increased collection of data and the widespread prevalence of algorithmic-selection applications (Ruckenstein & Granroth, 2020), further risks have been addressed. Key examples include worries about online *surveillance* based on the widespread tailoring of online contents (Ruckenstein & Granroth, 2020; Véliz, 2020; Zuboff, 2019), fears about *distorted information* through the algorithmic ranking of search results and news articles (Bozdag, 2013; Flaxman et al., 2016), and perceived *internet overuse* fostered by curated entertainment content by recommender systems (Gui & Büchi, 2019; Syvertsen, 2020). While concerns about the social risks of surveillance, distorted information, and internet overuse have existed before the spread of algorithmic-selection applications on the internet, the ubiquity of algorithms online renders their prevalence more severe; algorithmic selection significantly facilitates the unspecific collection and analysis of large amounts of personal data, the tailoring of contents based thereon as well as the allocation of personalized recommendations (Büchi et al., 2020).

Currently, statutory regulation (e.g., the General Data Protection Regulation (GDPR) in the European Union (EU)) as a governance mode to reduce such risks is increasing (see Larus et al., 2018). Despite this, a sense of helplessness and a wish

for more control over opaque algorithms remain prevalent sentiments among internet users (Festic, 2020). Self-help strategies—such as adjusting one's privacy settings—provide a complementary governance choice for internet users to cope with algorithmic risks (Boerman et al., 2018; Latzer & Just, 2020). Therefore, how users engage with algorithmic-selection applications and cope with their risks warrants attention (Kitchin, 2017; Ramizo, 2021).

This article provides previously lacking nationally representative data on how internet users cope with algorithmic risks. By doing so, this study contributes to a better understanding of factors that are associated with internet users' self-help strategies when coping with diverse algorithmic risks. Our theoretical basis for these mechanisms lies in three approaches that seek to explain how people react to risks and try to reduce them, i.e., the protection motivation theory (Rogers, 1975), the health belief model (Rosenstock, 1974), and the integrated behavior model (Montaño & Kasprzyk, 2008). Derived from these theories we propose that the awareness of a risk, the affectedness by this risk, and the skills related to the risk are associated with the use of self-help strategies against it. In line with this, this article seeks to show how these three factors are associated with internet users' self-help strategies against three types of algorithmic risks: *How are the awareness of algorithmic risks (1), the personal affectedness by these risks (2), and algorithm skills (3) associated with internet users' self-help to cope with the three algorithmic risks surveillance (S), distorted information (D), and internet overuse (O)?*

To investigate this question, we use structural equation modelling (SEM) on survey data representative for the Swiss online population to determine the association of three theory-derived factors with internet users' self-help strategies to cope with algorithmic risks. Our findings contribute to the discussion about how internet users' self-help strategies can be promoted as an alternate governance mode in an otherwise difficult to regulate space.

2 THEORETICAL BACKGROUND AND EXTANT RESEARCH ON ALGORITHMIC RISKS AND COPING STRATEGIES

2.1 Algorithmic Risks

Algorithmic selection relies on the constant and automated collection of massive amounts of data (Büchi et al., 2020; Hildebrandt, 2008), which entails a range of diverse risks, like feeling surveilled (Ruckenstein & Granroth, 2020; Zuboff, 2019), having one's privacy violated (Véliz, 2020), seeing filtered and personalized content on search engines and social media (Swart, 2021) that can be biased (Bozdag, 2013), distorted (Flaxman et al., 2016), and manipulative (Petre et al., 2019), and feeling like one is spending more time online than intended (Gui & Büchi, 2019; Syvertsen, 2020), which can affect internet users' well-being (Büchi et al., 2019).

Whenever internet users do something online, their behavior leaves data traces (Micheli et al., 2018; Ruckenstein & Granroth, 2020). On one hand, internet users can actively generate data by producing content, for instance by uploading an image to a social networking site. On the other hand, internet users can also (unconsciously) generate data when browsing the internet, for example when googling a certain product or clicking on an advertisement. These data traces can be collected, tracked, mined and evaluated algorithmically (Micheli et al., 2018), which leads to an increased risk of *surveillance*, for instance by platforms, governments, organizations or peers (Büchi et al., 2020; Demertzis et al., 2021; Zuboff, 2019). Furthermore, algorithmically selected content differs between individual internet users as it can be adapted according to their past behavior and interests (Bozdag, 2013; Gillespie, 2014; Swart, 2021). This personalization can lead to an increase in *distorted information*. In addition, the internet's relevance for various aspects of our everyday life together with receiving automated recommendations can lead to an increase in time spent online. Consequently, internet users can feel like they spend too much time online, which translates into *perceived internet overuse* (Büchi et al., 2019; Syvertsen, 2020). While previous research into algorithmic risks and the ways in which internet users cope with them have focused primarily on privacy protection (e.g., Boerman et al., 2018), we seek to extend this research by focusing on these three: surveillance, distorted information, and perceived internet overuse. One aspect that these risks have in common is that internet users can actively cope with them by engaging in dedicated self-help strategies. These self-help strategies are introduced in the following section.

2.2 Regulation of Algorithmic Risks: Self-Help Strategies

Reducing the algorithmically fueled risks introduced above is a goal of risk-based regulatory approaches (Latzer & Just, 2020). Such governance modes include statutory regulation (e.g., the GDPR in the EU), market solutions, and self-regulation of the industry (Latzer, Saurwein, et al., 2019; Latzer & Just, 2020; Saurwein et al., 2015; Seyfert, 2021). Despite statutory regulation aiming at increasing users' sovereignty over their own data, many internet users wish for more control over algorithms (Festic, 2020). One governance mode (Latzer & Festic, 2019; Latzer & Just, 2020), which complements statutory regulation and industry self-regulation, is self-management by users, for instance of their privacy (Boerman et al., 2018). We argue that applying such *self-help strategies* is a valid complementary governance choice for internet users to cope with risks that are related to algorithmic selection. The term 'self-help' originates in the domain of psychology. It designates the adaptation of one's own behavior to cope with problems and has spilled into other academic fields as well as popular culture (see Illouz, 2008 for a critical appraisal of the term). Therefore, we use this term to highlight individuals exerting agency when coping with algorithmic risks.

From a user perspective, there are many ways to deal with algorithmic risks. To mitigate the *risk of surveillance*, internet users can try to make their online habits less traceable (Büchi et al., 2017; Micheli et al., 2018; Sánchez & Viejo, 2018), for instance, by adjusting their privacy settings, using virtual private networks (VPNs) (Longworth, 2018; Weinberger et al., 2017), using their browser's private mode, deleting cookies (Boerman et al., 2018; Park, 2015), or applying privacy-enhancing technologies like the browser add-on Ghostery (Ireland, 2020; Latzer & Just, 2020). Moreover, internet users can use online content selectively or even refrain from using certain services (Boerman et al., 2018) and thereby, produce less data that can be used as input for algorithmic selection. Such strategies can be understood as preventive (Ebbers, 2020). To alleviate the *risk of distorted information*, users can double check information that they see online, for instance displayed on their social media news feeds (Islam et al., 2020; Leeder, 2019). Thereby, they can react to the content that has been algorithmically curated for them in a critical way (Zarouali et al., 2017). Such behaviors can be seen as defensive (Ebbers, 2020). To reduce the *risk of perceived internet overuse*, internet users can limit their screen-time or abstain from using certain services (at certain times) (Syvertsen, 2020), or ignore the automated recommendations that they are shown online. They can also try to influence the algorithmic content they see, for instance, by (not) liking or (not) clicking on certain content to inform the algorithm about their interests and preferences (Cotter, 2019; Gan, 2017; Lowe-Calverley & Grieve, 2018; Marder, 2018) or by (un-)following accounts or hiding certain posts in their timeline (Swart, 2021).

In sum, internet users can apply a variety of self-help strategies when interacting with algorithmic-selection applications to cope with the risks their use can entail. The degree to which self-help strategies pose an effective way to mitigate algorithmic risks remains difficult to estimate due to the black-box nature of algorithmic selection and the opacity of the services in which it is embedded (Kitchin, 2017). Still, taking action by applying such self-help strategies is a way in which internet users exert agency and regain autonomy in the digital space. As has been shown for privacy protection behavior (Büchi et al., 2021), the application of protective behavior is highly unequally distributed in digital societies. To understand who applies self-help strategies online to cope with algorithmic risks, the following section introduces a set of important factors to consider in this context.

2.3 Factors Associated with the Use of Self-Help Strategies

The theoretical approaches that build the basis for our model explaining how different factors influence how internet users cope with algorithmic risks originate in the realm of health protective behavior. These models were originally conceptualized to explain with what factors behaviors against health risks (e.g., smoking cessation, HIV-prevention) are associated. Recently, such approaches

have been transferred to the field of communication research to study protective behavior that reduces risks that internet use entails, like risks related to privacy protection or online behavioral advertising (e.g., Boerman et al., 2018; Ham, 2017). The use of these approaches brings the benefit of applying established theories on behavioral mechanisms to a new context. This article's hypotheses are rooted in three such theoretical approaches: the protection motivation theory (Rogers, 1975), the health belief model (Rosenstock, 1974), and the integrated behavior model (Montaño & Kasprzyk, 2008). Taken together, these approaches propose that whether and to what extent a person applies certain behaviors to reduce a specific type of risk depends on the perceived severity of this risk and the perceived personal susceptibility to it, as well as on a person's knowledge about and attitude towards it. We transfer these theoretical approaches that are geared towards explaining protective behavior against risks in more general terms to the field of algorithmic risks. Hence, we integrate these three theoretical approaches to explain what factors are associated with internet users' self-help strategies against algorithmic risks. For each of these factors, we will show how these theoretical approaches together with existing research led to our hypotheses.

2.3.1   *Risk Awareness and Self-Help Strategies*

To begin with, the protection motivation theory (Rogers, 1975) and the health belief model (Rosenstock, 1974) propose that the perceived severity of a risk influences whether someone intends to apply protective behaviors to reduce a risk. Findings about the relationship between concerns and protection strategies in the field of online risks differ according to type of protection measures that are applied. For instance, no or only a partial relation between privacy-related concerns and the (non-)use of social networking sites (Baruh et al., 2017) or smart speakers (Lutz & Newlands, 2021) has been found. At the same time, an association between privacy concerns and the general use of protection measures has been found by several survey studies (e.g., H. Chen et al., 2017; Dienlin & Metzger, 2016 for SNS; Ireland, 2020), including a meta-analysis of studies on privacy management (Baruh et al., 2017). A two-wave panel study (Boerman et al., 2018) that applied the protection motivation theory (Rogers, 1975) to privacy protection online indicates that firstly, people are aware of the data that is being collected about them and perceive this as problematic and secondly, the perceived severity of a privacy threatening risk, i.e., users' perception of its seriousness (Witte, 1992), is associated with their protective behavior. Based on this existing literature, we derived the following hypothesis for our study:

   H1: Risk awareness is positively associated with internet users' application of self-help strategies to cope with algorithmic risks.

2.3.2   Personal Risk Affectedness and Self-Help Strategies

Furthermore, protective behavior has been theorized to be associated with one's prior experience regarding a risk (Rogers, 1975) as well as the perceived susceptibility to it (Rosenstock, 1974). There is robust empirical evidence for this relationship regarding online risks: several empirical studies have shown that having experienced that one's privacy has been violated or feeling that it could be violated leads to increased levels of privacy protection and an increase in applying privacy-enhancing techniques and technologies when using social media (see Debatin et al., 2009) or the internet in general (see Büchi et al., 2017; H. Chen & Atkin, 2020; Ireland, 2020). Having experienced privacy breaches relates to an increased level of awareness of this risk (Baek et al., 2014; Cho et al., 2010). Deducted from these findings, we propose the following hypothesis:

H2: Personal risk affectedness is positively associated with internet users' application of self-help strategies to cope with algorithmic risks.

2.3.3   Algorithm Skills and Self-Help Strategies

Finally, another aspect that has been found to be central for individual risk protection behavior, is the perceived self-efficacy to cope with a risk (Rogers, 1975; Rosenstock, 1974) or a users' knowledge or skills related to this risk (Montaño & Kasprzyk, 2008). Congruently, findings from several representative survey studies focusing on internet use have shown that users' response-efficacy or self-efficacy is relevant for their protection behavior to reduce risks related to their privacy (Boerman et al., 2018; Dienlin & Metzger, 2016; Ham, 2017). In the same way, users' privacy literacy and internet skills have been shown to be associated with the degree to which they protect their privacy online (Bartsch & Dienlin, 2016; Baruh et al., 2017; Büchi et al., 2017; H. Chen & Atkin, 2020). More recently, besides traditional media literacy and internet skills (Hargittai, 2005; Litt, 2013; van Dijk, 2020), the omnipresence of algorithms in an online environment has led to a specific subset of internet skills coming into the focus of researchers. This specific type of internet skills relates to algorithmic selection and has been referred to as algorithm literacy or algorithm skills (see Dogruel et al., 2021; Gruber et al., 2021; Hargittai et al., 2020). Based on this extant research, we derived the following hypothesis:

H3: Algorithm skills are positively associated with internet users' application of self-help strategies to cope with algorithmic risks.

2.3.4   Path Model of Factors Associated with Self-Help Strategies

Based on the theoretical models and existing empirical research, the introduced hypotheses lead to the following path model of factors associated with self-help strategies to cope with algorithmic risks (see Figure 1).
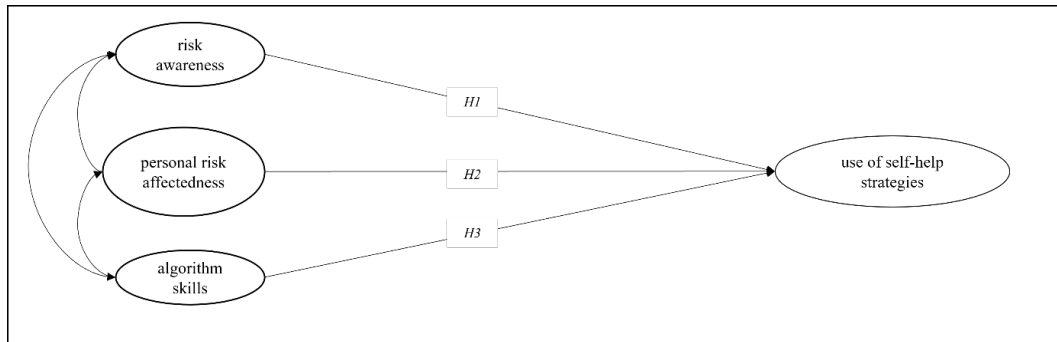
Figure 1. Path Model of Factors Associated with the Application of Self-Help Strategies to Cope with Algorithmic Risks. Source: Own illustration, based on Montano & Kasprzyk (2008), Rogers (1975), and Rosenstock (1974).

2.3.5   Interplay of Factors Associated with Self-Help Strategies

Regarding the relationships between these factors that are associated with self-help strategies to cope with algorithmic risks, we take the general theoretical approaches as a basis. Firstly, the protection motivation theory (Rogers, 1975) proposes that risk awareness correlates with personal risk affectedness and with skills. Secondly, the health belief model (Rosenstock, 1974) suggests that skills correlate with risk affectedness as well. With our model we apply these relationships to the field of algorithmic risks and hence to the awareness of algorithmic risks, the personal affectedness by algorithmic risks, and algorithm skills. Our model reflects these theoretical assumptions as the covariances between risk awareness, personal risk affectedness, and algorithm skills were estimated freely.

In addition to this model, sociodemographic background variables have been conceptualized to play a role as for people's application of protective behavior (Rosenstock, 1974). Previous research has shown that age, gender and the level of educational attainment are related to the awareness of algorithms and associated risks as well as to the level of skills and the application of protecting practices online (Cotter & Reisdorf, 2020; Park, 2011, 2015). At the same time, factors like one's experience with algorithms online (Cotter & Reisdorf, 2020; Swart, 2021) or a person's internet skills (Büchi et al., 2017) were found to be more important in explaining protective behavior than sociodemographic background variables.

2.4   Contributions

By empirically testing the theoretically derived model above, this article contributes to filling the following research gaps. While research on self-help strategies to cope with algorithmic risks is emerging, it has several blind spots. By mainly focusing on one specific application of algorithmic selection (e.g., personalized recommendations, curated social media news feed), or a certain population (e.g., users of one social networking site, youths), previous research offers limited insights

30

into the use of self-help strategies against algorithmic selection for general internet use. Moreover, most of the studies in the field have focused on single influencing factors on risk protection strategies. A more comprehensive analysis of influencing factors and their interplay is wanted. Furthermore, so far, research has predominantly focused on privacy protection practices (e.g., Boerman et al., 2018; Büchi et al., 2017; Ireland, 2020), although the list of risks associated with using algorithmic-selection applications is much more diverse. Further algorithmic risks like surveillance, distorted information, or perceived internet overuse have not been considered thoroughly yet, and accounts on the adoption of strategies to cope with such algorithmic risks are lacking so far. In addition, while recently, qualitative studies on the awareness of algorithms (Dogruel et al., 2020; Hargittai et al., 2020; Swart, 2021) or practices related to data collection (Selwyn & Pangrazio, 2018) were conducted, more generalizable findings and the systematic testing of possible associations are desired. Finally, many of the existing quantitative studies have been conducted in the US. Extending research beyond this context is essential for gaining relevant insights on a societal level. In sum, nationally representative, theory-driven and user-centric empirical studies on how internet users cope with diverse algorithmic risks and what factors play together in being associated with diverse self-help strategies are lacking. We aim to contribute to filling this gap by investigating how awareness of algorithmic risks, personal risk affectedness, and algorithm skills relate to the self-help strategies that users apply to cope with the algorithmic risks of surveillance, distorted information, and internet overuse. The following section describes the methodological design implemented to test the theoretical model introduced above.

3 METHOD

This section details the sample with which the survey was conducted, the measures used, as well as how the data was analyzed.

3.1 Sample

This article analyzes online survey data representative of Swiss internet users aged 16 and over (N_{2018}=1,202) regarding age, gender, household size, and employment status (see Table 1). The data was weighted to closely match the demographics in the general internet-user population. In Switzerland, at the time of data collection, 92% of the population used the internet (Latzer, Büchi, et al., 2019). The sample reflects the three biggest Swiss language regions. Data was collected between November 2018 and January 2019 by an independent market research company. All participants gave informed consent about their participation and the research design was approved by the university's ethics review board.

Table 1. Sample Characteristics. Note. N_{2018}=1,202; Swiss internet users aged 16 and over. Rounded percentages[1].

	Sample
Gender	
female	49%
male	51%
Age	
16-29 years	24%
30-44 years	28%
45-59 years	29%
60-79 years	19%
Education level	
low	7%
medium	67%
high	25%
Household income	
< 6,000 CHF	29%
> 6,000 CHF	71%

3.2 Measures

Central to our analysis are factors associated with the self-help strategies that internet users apply to cope with the algorithmic risks surveillance (*S*), distorted information (*D*) and internet overuse (*O*). Based on theoretical considerations and previous research we identified the following influencing factors on internet users' self-help strategies (see Figure 1): risk awareness (1), personal risk affectedness (2), and algorithm skills (3). For each type of risk, these concepts were measured differently, except for algorithm skills, which were measured consistently among risks[2].

Risk awareness. Respondents were asked how often they think about risks that are associated with algorithmic selection (1-4: *never – often*). These risks include for instance the constant monitoring of internet users (*S*), the danger of distorted information (*D*) or spending too much time online (*O*).

Personal risk affectedness. People were asked to what extent they feel personally affected by a list of online risks *(1-5: do not agree at all – strongly agree)*. This includes for instance feeling surveilled online (*S*), feeling confronted with untrue claims online (*D*) or thinking that they are relying too strongly on the internet (*O*).

---

[1] See federal statistical office https://www.bfs.admin.ch/asset/de/479-2000 for description of the Swiss population.
[2] See https://osf.io/c7aj3/?view_only=5e5343dce34e4486a1d0750642e1577f for exact wordings of all included items.

Algorithm skills. Respondents were asked to indicate their understanding of a list of terms related to the internet and algorithmic selection *(1-5: do not understand the term at all – completely understand the term).* This list included terms like 'algorithm' or 'personalized recommendation' that are related to the internet and algorithmic selection. Its design was adapted from Hargittai (2005), and the list was modified to reflect skills related to algorithms.

Self-help strategies. After having assessed the relevance of a list of risks that can be associated with using the internet, respondents were asked: "There are several things you can do to protect yourself or to deal with such risks. Please indicate how often *(1-5: never – always)* you do the following things". In this way, there were asked about the frequency with which they apply self-help strategies to cope with algorithmic risks. These strategies include adjusting one's privacy settings on certain online services as a strategy to reduce surveillance (S_{strat}), double-checking information online as a strategy to deal with distorted information (D_{strat}) and deliberately ignoring automated recommendations as a strategy to mitigate perceived internet overuse (O_{strat}).

Sociodemographic background. Respondents' gender (1=m, 2=f), age (1=16-29, 2=30-44, 3=45-59, 4=60-79), level of educational attainment (1=low, 2=medium, 3=high), and household income (1=< 6,000 CHF, 2 = > 6,000 CHF) were noted.

3.3 Data Analysis

We calculated a separate SEM using the package *lavaan* in *R* for each algorithmic risk, i.e., surveillance (*S*), distorted information (*D*), and overuse (*O*) to estimate the association of risk awareness (1), personal risk affectedness (2), and algorithm skills (3) with the application of self-help strategies for each of the algorithmic risks. As an estimator, we used maximum likelihood estimation. To deal with missing data, we used full information maximum likelihood estimation. All three measurement models showed an acceptable fit according to Hu and Bentler (1999): for surveillance, the fit indices were χ_S^2=95.977; df_S=24; p<.05; CFI_S=.963; TLI_S=.944; $RMSEA_S$=.050; $SRMR_S$=.034; for distorted information, the fit indices were χ_D^2=115.747; df_D=32; p<.05; CFI_D=.959; TLI_D=.942; $RMSEA_D$=.047; $SRMR_D$=.032; and for overuse, the fit indices were χ_O^2=43.719; df_O=21; p<.05; CFI_O=.974; TLI_O=.960; $RMSEA_O$=.047; $SRMR_O$=.020.

4   RESULTS

This section shows the frequency with which internet users apply the different self-help strategies and presents the results of the SEMs for each algorithmic risk.

Figure 2 depicts the results of the SEM for the algorithmic risk *surveillance (S)*, only displaying significant influencing paths (*p*<.05).

Figure 2. Factors Associated with Internet Users' Application of Self-Help Strategies to Cope with Surveillance. Note. Standardized path coefficients, p<.05; line width of hypothesized effects is scaled to the coefficients. N₂₀₁₈=1,202; Swiss internet users aged 16 and over.

For the risk of surveillance (S), the fit indices of the SEM were acceptable according to Hu and Bentler (1999): χ_S^2=113.407; df_S=30; p<.05; CFI_S=.960; TLI_S=.940; $RMSEA_S$=.048; $SRMR_S$=.033. 31% of internet users say that they adjust their privacy settings for certain internet services often or always[3]. The results reveal that this self-help strategy to cope with surveillance is positively associated with risk awareness, with personal risk affectedness and with the level of algorithm skills. Thus, for surveillance, we can accept hypotheses $H1_S$, $H2_S$ and $H3_S$. Furthermore, the covariances of all influencing factors were significant and positive, which is in line with our assumptions introduced above.

Figure 3 depicts the results of the SEM for the algorithmic risk *distorted information (D)*, only displaying significant influencing paths (*p*<.05).

---

[3] See https://osf.io/c7aj3/?view_only=5e5343dce34e4486a1d0750642e1577f for distribution of frequencies for all self-help strategies.

Figure 3. Factors Associated with Internet Users' Application of Self-Help Strategies to Cope with Distorted Information. Note. Standardized path coefficients, p<.05; line width of hypothesized effects is scaled to the coefficients. N$_{2018}$=1,202; Swiss internet users aged 16 and over.

For the risk of distorted information (*D*), the fit indices of the SEM were also acceptable: $\chi_D 2$=134.552; df_D=39; *p*<.05; CFI_D=.957; TLI_D=.939; $RMSEA_D$=.045; $SRMR_D$=.032. 46% of internet users state that they often or always double check information online by using additional information sources or different search engines. This self-help strategy to cope with distorted information is positively associated with risk awareness and with the level of algorithm skills, but not with risk affectedness. For the risk of distorted information, we can thus accept $H1_D$ and $H3_D$, but not $H2_D$. Again, the covariances between the influencing factors were significant and positive.

Figure 4 depicts the results of the SEM for the algorithmic risk *internet overuse (O)*, only displaying significant influencing paths (*p*<.05).
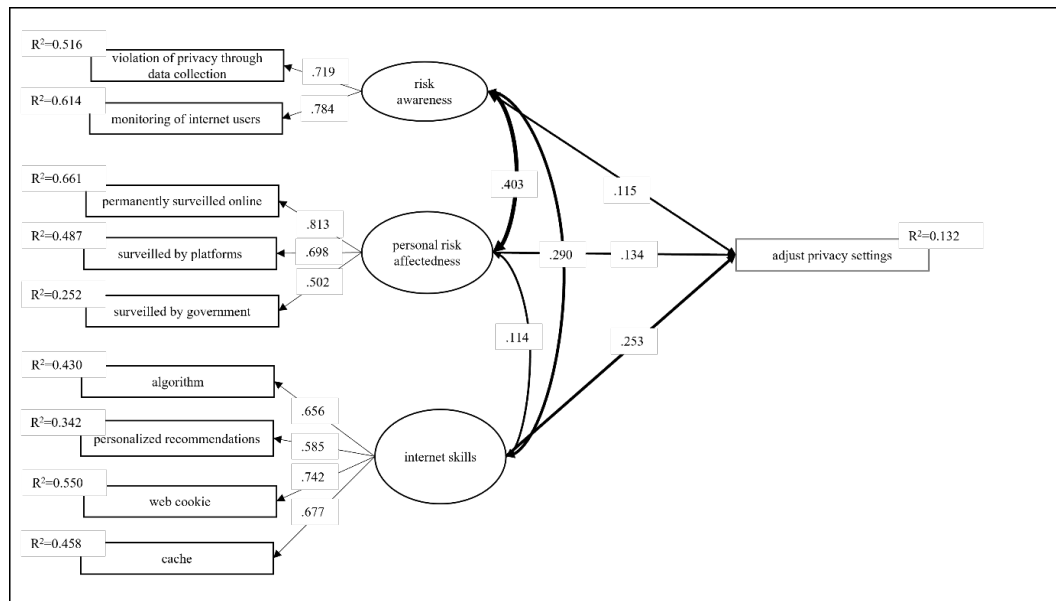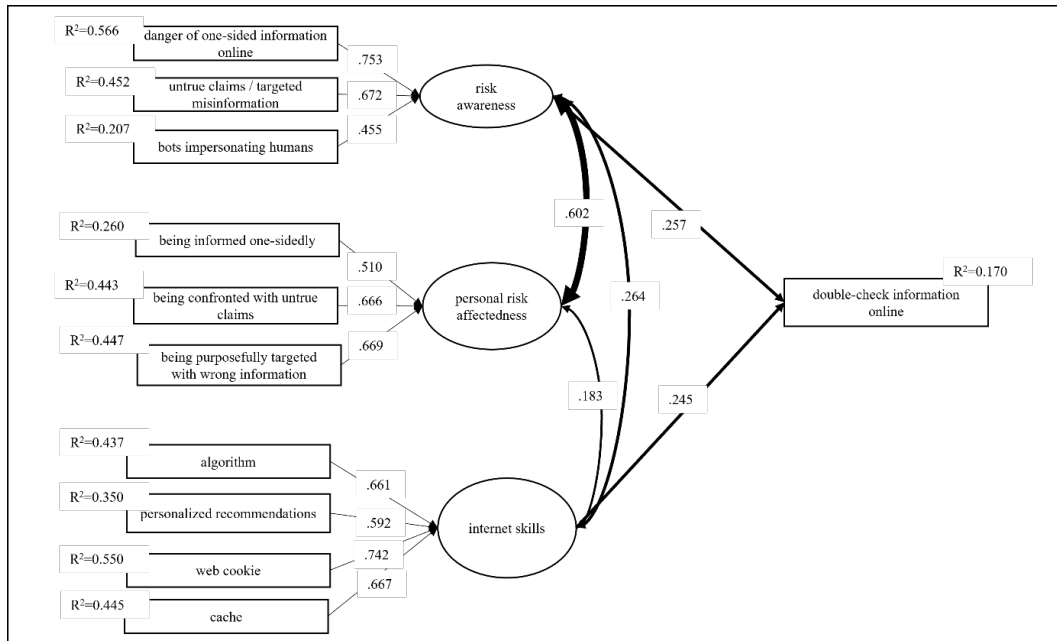
Figure 4. Factors Associated with Internet Users' Application of Self-Help Strategies to Cope with Internet Overuse. Note. Standardized path coefficients, p<.05; line width of hypothesized effects is scaled to the coefficients. N_{2018}=1,202; Swiss internet users aged 16 and over.

For the risk of internet overuse (O), the fit indices of the SEM were good as well: χ_O^2=52.389; df_O=16; $p<.05$; CFI_O=.972; TLI_O=.951; $RMSEA_O$=.044; $SRMR_O$=.020. 71% of internet users say that they often or always ignore automated recommendations. This self-help strategy to cope with internet overuse is positively associated with users' risk awareness and with their algorithm skills. Moreover, ignoring automated recommendations is negatively associated with personal risk affectedness. Hence, for overuse, $H1_O$ and $H3_O$ can be accepted. For $H2_O$ the direction of the effect is opposite to our expectations. In addition, awareness of perceived overuse covaried significantly positively with risk affectedness. There was no significant covariance between algorithm skills and either of the influencing factors.

Alternative models tested the effects of sociodemographic background variables (i.e., gender, age, educational attainment, and income) on risk awareness, personal risk affectedness, and algorithm skills. There were no notable differences in terms of effect sizes and directions as well as significance levels for the hypothesized associations when sociodemographic variables were included, indicating robustness of the models introduced above (Figures 2–4). However, the alternative models' fit was not satisfactory. Therefore, we decided to exclude sociodemographic background variables from our analysis to ameliorate our models' fit[4].

---

[4] Refer to https://osf.io/c7aj3/?view_only=5e5343dce34e4486a1d0750642e1577f for an overview over the SEMs including sociodemographic background variables.

5   DISCUSSION

The analyses of the SEMs highlight the importance of a higher level of awareness of algorithmic risks (1) and algorithm-specific algorithm skills (3) on the extent to which internet users apply self-help strategies to cope with the algorithmic risks surveillance (*S*), distorted information (*D*), and overuse (*O*). In addition, we found that personal risk affectedness (2) is positively associated with applying self-help strategies to cope with surveillance (*S*), negatively associated with self-help to cope with overuse (*O*), but not associated with self-help to cope with distorted information (*D*). The negative effect for overuse can possibly be explained through an exposure effect, i.e., a desire or need to use the internet extensively (e.g. through professional or private social pressure, see Gui & Büchi, 2019) may lead to experiencing personal risk affectedness like spending too much time on apps due to automated recommendations merely because a lot of time is spent online; needing to use the internet despite this and thus refraining from ignoring automated recommendations is not an unlikely behavior and in line with extant research on perceptions of internet overuse (Büchi et al., 2019).

Moreover, we found that awareness of algorithmic risks (1), personal risk affectedness (2), and algorithm skills (3) have significant and positive covariances with each other in the models for surveillance (*S*) and distorted information (*D*), but not for internet overuse (*O*).

At the same time, the frequency with which internet users apply different self-help strategies varies between types of risks. Overall, only few internet users apply self-help strategies against surveillance and distorted information on a regular basis. Previous research has suggested this for strategies to protect one's privacy in a similar way (see Boerman et al., 2018). On the other hand, ignoring automated recommendations is more widespread. Explanations for this variation can lie in different aspects of internet use. Viewing privacy as contextual integrity (Nissenbaum, 2010) highlights that users' judgement of data being shared differs with regard to context, actors, attributes and transmission principles (Vitak & Zimmer, 2020). Hence, internet users might judge certain self-help strategies regarding specific algorithmic risks more important than others which may lead to a difference in the use of self-help strategies.

Similarly, qualitative research has shown that the awareness of algorithms varies across different services (Swart, 2021). This may relate to the awareness of algorithmic risks and the felt need for applying strategies to cope with them. In addition, the motives for the use of certain services might be associated with users' online behavior. For instance, wanting to or having to use certain algorithmic applications like social media can override the wish for privacy (Quinn, 2016). Moreover, research has indicated that practices that mitigate the effect of algorithmic selection are often deemed too laborious by users (Kormelink Groot & Meijer Costera, 2014; Monzer et al., 2020). This can lead to users not taking advantage of such strategies even if they wished for more agency over the contents

that they are shown (Swart, 2021). Thus, the simpleness of use can be an important factor for self-help strategies, just like knowing how to implement such protective practices is (see Büchi et al., 2017). Another reason for not adopting a certain self-help strategy may lie in the habitual use of algorithmic applications. Such habits can be related to the behavior that one engages in (Montaño & Kasprzyk, 2008). For instance, social media users may not consider unfollowing accounts that they are no longer interested in (Swart, 2021). Social media applications are often woven into users' routines so that discontinuation of use would have severe consequences (Dienlin & Metzger, 2016). The different self-help strategies vary in their impact on users' daily internet use. For instance, not using a certain service at all has a different effect on internet use and its consequences than deleting cookies or ignoring recommendations. This may affect the use of such strategies as well, and consequently, users may refrain from applying them in the first place. In addition, research has indicated that internet users do not see the automated data collection and algorithmic analysis thereof as problematic in the first place, as they state that they have nothing to hide (Demertzis et al., 2021). This suggests that in addition to the possibility that individuals are not aware of risks or do not feel affected by risks, they can also take on an attitude of having nothing to hide and thereby not feel a need to apply any strategies to counter possible risks.

Specifically in Switzerland, people may feel rather certain about digital risks as the GDPR from the EU is applied by many corporations that are operating in Switzerland as well. Recent research has shown that the regulatory context of a country can play a role for internet users' felt need to change how they behave because of potential online harms (Strycharz et al., 2022). At the same time, research in Switzerland has shown that only 25% of internet users do not feel exposed to any dangers when they are online. This suggests that there is still some general skepticism towards safety online (Latzer et al., 2021).

Finally, the use of self-help strategies illustrates that while algorithmic-selection applications exert power over their users, the users also have agency to use those platforms to their ends by acting strategically (Bakardjieva, 2005; Ramizo, 2021; Selwyn & Pangrazio, 2018; van Dijck, 2009). The interaction with algorithmic-selection applications can in turn influence the algorithms as in an online environment, humans and algorithms form a recursive loop (Bucher, 2017; Gillespie, 2014). However, users seem to not always be aware of this reciprocated relationship (Swart, 2021). Assigning the responsibility for data protection in algorithmic environments fully to the users is therefore problematic (Baruh & Popescu, 2017).

The theoretical basis of this study roots in psychological concepts originating in the field of health protection behavior (Montaño & Kasprzyk, 2008; Rogers, 1975; Rosenstock, 1974), to derive how users cope with risks online. In the field of privacy protection, the privacy paradox (e.g., H.-T. Chen, 2018; Gerber et al., 2018) and privacy calculus theory (e.g., H.-T. Chen, 2018; Dienlin & Metzger, 2016; Gutierrez et al., 2019) are approaches that try to explain why internet users

engage with social media despite potential privacy-related risks. The mechanisms that these theories propose may also inform the analysis of further algorithmic applications that entail similar risks, and therefore, these concepts could be incorporated in future research on algorithmic risk protection in a broader sense.

For this study, there are a few limitations to consider. First, our study includes a variety of theoretically derived and empirically identified factors that are associated with internet users' self-help strategies aimed at reducing algorithmic risks. Besides the factors that we identified based on our theoretical approaches, further factors that could be associated with internet users' self-help strategies are imaginable. For instance, previous research suggests that age, education, and gender may affect online privacy protection (Büchi et al., 2021). In addition, future research could focus on deriving additional potential influencing factors. For example, internet users' actions can be related to their trust in certain websites and services (Pengnate & Sarathy, 2017). The degree of transparency of algorithmic processes may be associated with users' behavior as well (Dogruel et al., 2020; Kemper & Kolkman, 2019). Moreover, besides the described factors, external shocks (Rosenstock, 1974), like privacy scandals made public in the media, may also play a role on the extent to which users apply self-help strategies to protect themselves (Büchi et al., 2022). At the same time, Swart's (2021) qualitative interviews indicate that such scandals can be common knowledge among social media users without leading them to stop using a certain service. Second, we decided to look at three specific risks that relate to algorithmic selection: surveillance, distorted information, and internet overuse. Future studies could include more digital risks, like for instance discrimination through algorithmic selection (Noble, 2018). Third, we looked at self-help strategies against algorithmic risks in general. In the future, such self-help strategies could be investigated in their relative context, for instance with case studies on self-help strategies regarding algorithmic selection on specific platforms like Instagram or context-specific privacy behavior for instance related to online purchases.

6   CONCLUSION

This study identifies factors that are associated with internet users' self-help strategies to cope with algorithmic risks. We found that internet users adjust their privacy settings, double-check information, and ignore automated recommendations to cope with the algorithmic risks of surveillance, distorted information, and internet overuse to varying degrees. The empirical results from our study representative of Swiss internet users showed that their risk awareness (1), their personal risk affectedness (2), and their level of algorithm skills (3) are important influencing factors on internet users' self-help strategies to cope with these algorithmic risks.

Self-help strategies are a valuable mechanism for the reduction of algorithmic risks. They provide a complementary governance option to the existing and

emerging governance mix composed of regulation that is issued by state authorities, industry's self-regulation and market solutions. Appropriate statutory regulations and clear guidelines are a prerequisite for the successful and adequate implementation of such complementary self-help strategies that internet users can apply.

This study analyzed the association of three theoretically derived factors with self-help strategies to cope with three types of algorithmic risks that concern internet users in their everyday digital life. It extends extant research about algorithmic risks mostly limited to threats to privacy and contributes to the field of governance of algorithms more broadly. Thereby, it provides an empirical basis for deducting the apt governance mix and assessing the role that users' self-help could play therein to cope with algorithmic risks.

FUNDING STATEMENT AND ACKNOWLEDGMENTS

REFERENCES

Baek, Y. M., Kim, E., & Bae, Y. (2014). My privacy is okay, but theirs is endangered: Why comparative optimism matters in online privacy concerns. *Computers in Human Behavior*, *31*, 48–56. https://doi.org/10.1016/j.chb.2013.10.010

Bakardjieva, M. (2005). *Internet Society: The Internet in everyday life.* Sage. http://dx.doi.org/10.4135/9781446215616

Bartsch, M., & Dienlin, T. (2016). Control your Facebook: An analysis of online privacy literacy. *Computers in Human Behavior*, *56*, 147–154. https://doi.org/10.1016/j.chb.2015.11.022

Baruh, L., & Popescu, M. (2017). Big data analytics and the limits of privacy self-management. *New Media & Society*, *19*(4), 579–596. https://doi.org/10.1177/1461444815614001

Baruh, L., Secinti, E., & Cemalcilar, Z. (2017). Online privacy concerns and privacy management: A meta-analytical review. *Journal of Communication*, *67*(1), 26–53. https://doi.org/10.1111/jcom.12276

Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2018). Exploring motivations for online privacy protection behavior: Insights from panel data. *Communication Research*, 1–25. https://doi.org/10.1177/0093650218800915

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227. https://doi.org/10.1007/s10676-013-9321-6

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Büchi, M., Festic, N., Just, N., & Latzer, M. (2021). Digital inequalities in online privacy protection: Effects of age, education and gender. *Handbook of Digital Inequality*. https://www.elgaronline.com/view/edcoll/9781788116565/9781788116565.00029.xml

Büchi, M., Festic, N., & Latzer, M. (2019). Digital overuse and subjective well-being in a digitized society. *Social Media + Society*. https://doi.org/10.1177/2056305119886031

Büchi, M., Festic, N., & Latzer, M. (2022). The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. *Big Data & Society*, *9*(1), 1–14. https://doi.org/10.1177/20539517211065368

Büchi, M., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A., Velidi, S., & Viljoen, S. (2020). The chilling effects of algorithmic profiling: Mapping the issues. *Computer Law & Security Review*, *36*, 1–15. https://doi.org/10.1016/j.clsr.2019.105367

Büchi, M., Just, N., & Latzer, M. (2017). Caring is not enough: The importance of Internet skills for online privacy protection. *Information, Communication & Society*, *20*(8), 1261–1278. https://doi.org/10.1080/1369118X.2016.1229001

Chen, H., & Atkin, D. (2020). Understanding third-person perception about Internet privacy risks. *New Media & Society*, 1–19. https://doi.org/10.1016/j.chb.2017.01.003

Chen, H., Beaudoin, C. E., & Hong, T. (2017). Securing online privacy: An empirical test on Internet scam victimization, online privacy concerns, and privacy protection behaviors. *Computers in Human Behavior*, *70*, 291–302. https://doi.org/10.1016/j.chb.2017.01.003

Chen, H.-T. (2018). Revisiting the privacy paradox on social media with an extended privacy calculus model: The effect of privacy concerns, privacy self-efficacy, and social capital on privacy management. *American Behavioral Scientist*, *62*(10), 1392–1412. https://doi.org/10.1177/0002764218792691

Cho, H., Lee, J.-S., & Chung, S. (2010). Optimistic bias about online privacy risks: Testing the moderating effects of perceived controllability and prior experience. *Computers in Human Behavior*, *26*(5), 987–995. https://doi.org/10.1016/j.chb.2010.02.012

Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, *21*(4), 895–913. https://doi.org/10.1177/1461444818815684

Cotter, K., & Reisdorf, B. C. (2020). Algorithmic knowledge gaps: A new horizon of (digital) inequality. *International Journal of Communication*, *14*(0), 21.

Debatin, B., Lovejoy, J. P., Horn, A.-K., & Hughes, B. N. (2009). Facebook and online privacy: attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, *15*(1), 83–108. https://doi.org/10.1111/j.1083-6101.2009.01494.x

Demertzis, N., Mandenaki, K., & Tsekeris, C. (2021). Privacy attitudes and behaviors in the age of post-privacy: An empirical approach. *Journal of Digital Social Research*, *3*(1), 119-152-119–152. https://doi.org/10.33621/jdsr.v3i1.75

Dienlin, T., & Metzger, M. J. (2016). An extended privacy calculus model for SNSs: Analyzing self-disclosure and self-withdrawal in a representative U.S. sample. *Journal of Computer-Mediated Communication*, *21*(5), 368–383. https://doi.org/10.1111/jcc4.12163

Dogruel, L., Facciorusso, D., & Stark, B. (2020). 'I'm still the master of the machine.' Internet users' awareness of algorithmic decision-making and their perception of its effect on their autonomy. *Information, Communication & Society*, 1–22. https://doi.org/10.1080/1369118X.2020.1863999

Dogruel, L., Masur, P., & Joeckel, S. (2021). Development and validation of an algorithm literacy scale for internet users. *Communication Methods and Measures*, 1–19. https://doi.org/10.1080/19312458.2021.1968361

Ebbers, F. (2020). How to protect my privacy? Classifying end-user information privacy protection behaviors. In *Privacy and Identity Management. Data for Better Living: AI and Privacy* (pp. 327–342). https://doi.org/10.1007/978-3-030-42504-3_21

Festic, N. (2020). Same, same, but different! Qualitative evidence on how algorithmic selection applications govern different life domains. *Regulation & Governance*, rego.12333. https://doi.org/10.1111/rego.12333

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. https://doi.org/10.1093/poq/nfw006

Gan, C. (2017). Understanding WeChat users' liking behavior: An empirical study in China. *Computers in Human Behavior*, *68*, 30–39. https://doi.org/10.1016/j.chb.2016.11.002

Gerber, N., Gerber, P., & Volkamer, M. (2018). Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, *77*, 226–261. https://doi.org/10.1016/j.cose.2018.04.002

Gillespie, T. (2014). *The relevance of algorithms. Media Technologies.* https://doi.org/10.7551/mitpress/9780262525374.003.0009

Gruber, J., Hargittai, E., Karaoglu, G., & Brombach, L. (2021). Algorithm awareness as an important internet skill: The case of voice assistants. *International Journal of Communication*, *15*(0), 19.

Gui, M., & Büchi, M. (2019). From use to overuse: Digital inequality in the age of communication abundance. *Social Science Computer Review*, 089443931985116. https://doi.org/10.1177/0894439319851163

Gutierrez, A., O'Leary, S., Rana, N. P., Dwivedi, Y. K., & Calle, T. (2019). Using privacy calculus theory to explore entrepreneurial directions in mobile location-based advertising: Identifying intrusiveness as the critical risk factor. *Computers in Human Behavior*, *95*, 295–306. https://doi.org/10.1016/j.chb.2018.09.015

Ham, C.-D. (2017). Exploring how consumers cope with online behavioral advertising. *International Journal of Advertising*, *36*(4), 632–658. https://doi.org/10.1080/02650487.2016.1239878

Hargittai, E. (2005). Survey measures of web-oriented digital literacy: *Social Science Computer Review*. https://doi.org/10.1177/0894439305275911

Hargittai, E., Gruber, J., Djukaric, T., Fuchs, J., & Brombach, L. (2020). Black box measures? How to study people's algorithm skills. *Information, Communication & Society*, *23*(5), 764–775. https://doi.org/10.1080/1369118X.2020.1713846

Hildebrandt, M. (2008). Defining profiling: A new type of knowledge? In M. Hildebrandt & S. Gutwirth (Eds.), *Profiling the European citizen: Cross-disciplinary perspectives* (pp. 17–45). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6914-7_2

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Illouz, E. (2008). *Saving the modern soul: Therapy, Emotions, and the culture of self-help*. University of California Press. https://www.jstor.org/stable/10.1525/j.ctt1pp4br

Ireland, L. (2020). Predicting online target hardening behaviors: An extension of routine activity theory for privacy-enhancing technologies and techniques. *Deviant Behavior*, 1–17. https://doi.org/10.1080/01639625.2020.1760418

Islam, A. K. M. N., Laato, S., Talukder, S., & 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. https://doi.org/10.1016/j.techfore.2020.120201

Just, N., & Latzer, M. (2017). Governance by algorithms: Reality construction by algorithmic selection on the Internet. *Media, Culture & Society*, *39*(2), 238–258. https://doi.org/10.1177/0163443716643157

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society, 22*(14), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Kormelink Groot, T., & Meijer Costera, I. (2014). Tailor-made news. *Journalism Studies, 15*(5), 632–641. https://doi.org/10.1080/1461670X.2014.894367

Larus, J., Hankin, C., Carson, S. G., Christen, M., Crafa, S., Grau, O., Kirchner, C., Knowles, B., McGettrick, A., Tamburri, D. A., & Werthner, H. (2018). *When computers decide: European recommendations on machine-learned automated decision making* [Technical Report]. Association for Computing Machinery.

Latzer, M., Festic, N., & Kappeler, K. (2020). Awareness of risks related to algorithmic selection in Switzerland. Report 3 from the project: *The significance of algorithmic selection for everyday life: The case of Switzerland.* Zurich: University of Zurich. http://mediachange.ch/research/algosig

Latzer, M., Büchi, M., & Festic, N. (2019). *Internetverbreitung und digitale Bruchlinien in der Schweiz 2019. Themenbericht aus dem World Internet Project—Switzerland 2019.* Universität Zürich. http://mediachange.ch/research/wip-ch-2019

Latzer, M., & Festic, N. (2019). A guideline for understanding and measuring algorithmic governance in everyday life. *Internet Policy Review, 8*(2). https://doi.org/10.14763/2019.2.1415

Latzer, M., Hollnbuchner, K., Just, N., & Saurwein, F. (2016). The economics of algorithmic selection on the internet. *Handbook on the economics of the internet.* https://www.elgaronline.com/view/edcoll/9780857939845/9780857939845.00028.xml

Latzer, M., & Just, N. (2020). Governance by and of algorithms on the internet: Impact and consequences. In M. Latzer & N. Just, *Oxford research encyclopedia of communication.* Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.904

Latzer, M., Saurwein, F., & Just, N. (2019). Assessing policy II: Governance-choice method. In H. Van den Bulck, M. Puppis, K. Donders, & L. Van Audenhove (Eds.), *The Palgrave handbook of methods for media policy research* (pp. 557–574). Springer International Publishing. https://doi.org/10.1007/978-3-030-16065-4_32

Leeder, C. (2019). How college students evaluate and share "fake news" stories. *Library & Information Science Research, 41*(3), 100967. https://doi.org/10.1016/j.lisr.2019.100967

Litt, E. (2013). Measuring users' internet skills: A review of past assessments and a look toward the future. *New Media & Society*, *15*(4), 612–630. https://doi.org/10.1177/1461444813475424

Longworth, J. (2018). VPN: From an obscure network to a widespread solution. *Computer Fraud & Security*, *2018*(4), 14–15. https://doi.org/10.1016/S1361-3723(18)30034-4

Lowe-Calverley, E., & Grieve, R. (2018). Thumbs up: A thematic analysis of image-based posting and liking behaviour on social media. *Telematics and Informatics*, *35*(7), 1900–1913. https://doi.org/10.1016/j.tele.2018.06.003

Lutz, C., & Newlands, G. (2021). Privacy and smart speakers: A multi-dimensional approach. *The Information Society*, *0*(0), 1–16. https://doi.org/10.1080/01972243.2021.1897914

Marder, B. (2018). Trumped by context collapse: Examination of 'Liking' political candidates in the presence of audience diversity. *Computers in Human Behavior*, *79*, 169–180. https://doi.org/10.1016/j.chb.2017.10.025

Micheli, M., Lutz, C., & Büchi, M. (2018). Digital footprints: An emerging dimension of digital inequality. *Journal of Information, Communication and Ethics in Society*, *16*(3), 242–251. https://doi.org/10.1108/JICES-02-2018-0014

Montaño, D. E., & Kasprzyk, D. (2008). Theory of reasoned action, theory of planned behavior, and the integrated behavioral model. In K. Glanz, B. K. Rimer & K, Viswanath (Eds.), *Health behavior and health education. Theory, research, and practice*. Jossey-Bass.

Monzer, C., Moeller, J., Helberger, N., & Eskens, S. (2020). User perspectives on the news personalisation process: Agency, trust and utility as building blocks. *Digital Journalism*, *8*(9), 1142–1162. https://doi.org/10.1080/21670811.2020.1773291

Nissenbaum, H. F. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford Law Books.

Noble, S. U. (2018). *Algorithms of oppression.* De Gruyter. https://doi.org/10.18574/nyu/9781479833641.001.0001

Park, Y. J. (2011). Digital literacy and privacy behavior online. *Communication Research, 4*(2), 215–236. https://doi.org/10.1177/0093650211418338

Park, Y. J. (2015). Do men and women differ in privacy? Gendered privacy and (in)equality in the Internet. *Computers in Human Behavior*, *50*, 252–258. https://doi.org/10.1016/j.chb.2015.04.011

Pengnate, S. (Fone), & Sarathy, R. (2017). An experimental investigation of the influence of website emotional design features on trust in unfamiliar online vendors. *Computers in Human Behavior*, *67*, 49–60. https://doi.org/10.1016/j.chb.2016.10.018

Petre, C., Duffy, B. E., & Hund, E. (2019). "Gaming the System": Platform paternalism and the politics of algorithmic visibility. *Social Media + Society*, *5*(4), 1–12. https://doi.org/10.1177/2056305119879995

Quinn, K. (2016). Why we share: A uses and gratifications approach to privacy regulation in social media use. *Journal of Broadcasting & Electronic Media*, *60*(1), 61–86. https://doi.org/10.1080/08838151.2015.1127245

Ramizo, G. J. (2021). Platform playbook: A typology of consumer strategies against algorithmic control in digital platforms. *Information, Communication & Society*, 1–16. https://doi.org/10.1080/1369118X.2021.1897151

Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change. *The Journal of Psychology*, *91*(1), 93–114. https://doi.org/10.1080/00223980.1975.9915803

Rosenstock, I. M. (1974). Historical origins of the health belief model. *Health education monographs*. https://doi.org/10.1177/109019817400200403

Ruckenstein, M., & Granroth, J. (2020). Algorithms, advertising and the intimacy of surveillance. *Journal of Cultural Economy*, *13*(1), 12–24. https://doi.org/10.1080/17530350.2019.1574866

Sánchez, D., & Viejo, A. (2018). Privacy-preserving and advertising-friendly web surfing. *Computer Communications*, *130*, 113–123. https://doi.org/10.1016/j.comcom.2018.09.002

Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: Options and limitations. *Info*, *17*(6), 35–49. https://doi.org/10.1108/info-05-2015-0025

Selwyn, N., & Pangrazio, L. (2018). Doing data differently? Developing personal data tactics and strategies amongst young mobile media users. *Big Data & Society*, *5*(1), 2053951718765021. https://doi.org/10.1177/2053951718765021

Seyfert, R. (2021). Algorithms as regulatory objects. *Information, Communication & Society*, 1–17. https://doi.org/10.1080/1369118X.2021.1874035

Strycharz, J., Kim, E. & Segijn, C. B. (2022). Why people would (not) change their media use in response to perceived corporate surveillance. *Telematics and Informatics, 71*. https://doi.org/10.1016/j.tele.2022.101838

Swart, J. (2021). Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. *Social Media + Society*, *7*(2), 20563051211008828. https://doi.org/10.1177/20563051211008828

Syvertsen, T. (2020). *Digital detox*. Emerald Publishing Limited. https://books.emeraldinsight.com/page/detail/Digital-Detox/?k=9781787693425

van Dijck, J. (2009). Users like you? Theorizing agency in user-generated content. *Media, Culture & Society*, *31*(1), 41–58. https://doi.org/10.1177/0163443708098245

van Dijk, J. (2020). *The digital divide*. Polity.

Véliz, C. (2020). *Privacy is power*. Bantom Press. /books/1120394/privacy-is-power/9780552177719

Vitak, J., & Zimmer, M. (2020). More than just privacy: Using contextual integrity to evaluate the long-term risks from COVID-19 surveillance technologies. *Social Media + Society*, *6*(3),1–4. https://doi.org/10.1177/2056305120948250

Weinberger, M., Bouhnik, D., & Zhitomirsky-Geffet, M. (2017). Factors affecting students' privacy paradox and privacy protection behavior. *Open Information Science*, *1*(1). https://doi.org/10.1515/opis-2017-0002

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs*, *59*(4), 329–349. https://doi.org/10.1080/03637759209376276

Zarouali, B., Ponnet, K., Walrave, M., & Poels, K. (2017). "Do you like cookies?" Adolescents' skeptical processing of retargeted Facebook-ads and the moderating role of privacy concern and a textual debriefing. *Computers in Human Behavior*, *69*, 157–165. https://doi.org/10.1016/j.chb.2016.11.050

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

AI AS SOCIAL ACTOR: A LACANIAN INVESTIGATION INTO SOCIAL TECHNOLOGY

Marc Heimann[a] and Anne-Friederike Hübener[a]

ABSTRACT

Given the social and political influence of social networks, which are often structured and organized by what today falls under the umbrella term *artificial intelligence*, we seek to define this new social frame. Most importantly, we ask how to frame this new social sphere in current theory and how it can be conceptualized for social sciences. However, this is not possible without constructing a logical frame for a problem as deeply entwined with the modern history of logic as AI is. We will therefore frame the problem of AIs as social actors within the logical discourse that Lacanian psychoanalysis opened. Our analysis shows that the inherent indeterminate that constitutes the psychoanalytic subject is omitted from AI-supplanted identities. Logical analysis also allows us to discern a specific mode of subjectivation that is made much more prominent through the normalization of phenomena like echo chambers and online identities.

Keywords: Artificial Intelligence, Philosophy, Psychoanalysis, Logic, Social Actors.

[a] Hochschule Niederrhein, Germany

1   INTRODUCTION

Within the Fourth Industrial Revolution, artificial intelligence (AI) has become a central element of our communication and social reality. But the question remains of how to approach this new element of social life. Is it really just "new wine in old wineskins?" (Pence, 2019), or does the change in the social landscape indicate the emergence of something new? Do we need, for example, tech social workers (Mathiyazhagan, 2021) as a result of these changes? Framing the social dimensions of technology as a tool or a prosthesis, as Freud did (Freud, 1968: 451), seems bound to fail. AI is today also a socially active artifact as it frames and organizes social connections and creates social identities. This also means that it interacts with the very nature of how we understand ourselves.

What do we understand under the umbrella term of artificial intelligence here? Most importantly those uses of machine learning and deep learning, which today are used to create new social bonds. Facebook for example expressively uses machine learning in layers to generate their social feeds (Lada et al., 2021). However, the following discussion will not focus on a specific model, but on the foundations of applied computer science in logic and applied mathematics. Now, a negative is hard to prove, but there are strong indications that current 'thinking machines' and artificial thinking in general has hit a limit in applicability to social questions. While it has emerged as an important tool for many sciences, it has also failed so far to discern any structural rules within social realities despite massive scientific effort (Dressel and Farid, 2018; Littlefield et al., 2021; Salganik et al., 2020). All these studies, of which the Salganik et al. should be put into focus, show the low predictive power of current AI models with regard to social data. This, understood as a "proximate failure" (Harman, 2016: 106), i.e., an inherent deadlock of it as a social object, is of major relevance to AI as a social actor. Because despite its inability to understand social structures, there is no indication that artificial curation of social connections will cease. To approach this proximate failure, however, a more abstract way of approaching the problem is needed. The reason for this is simple: if, as the study by Salganik et al. shows, a broad range of AI models fail to produce any substantial knowledge, it makes no sense to analyze specific cases of AI failure. Instead, it should be assumed that there is a more fundamental problem at hand. This is why the focus will be on the social ramifications of a particular understanding of logic, which organizes and structures social links created and curated by machines.

2   SOCIAL ACTION

Previously, the constitution of social identities has been thought of as being based in shared experiences, thereby bridging the problem of representation in socio-linguistic concepts of social identity. However, this shared experience is increasingly supplanted by the technological link that social algorithms provide. What is this

supplantation? Everyone who uses social media today is aware of their content feed being an algorithmically curated (often with specific economic intent) display of other user-created content and advertising. If we are connected with other people through social media, which uses machine-based curation of content, be it YouTube, Facebook or others, this linkage itself is curated, not just the appearance of it on our interfaces. This means that AI's, in a broad sense of automated judgment through computing, but also as machine learning and deep learning in the case of Meta, acts directly on the interlinking of social actors, without appearing as a social actor itself. Instead of essentially random face-to-face experiences in which we meet people, the AI curating the feed of a Facebook user supplants this experience, by replacing the random, partially virtual and complex character of this with a positive mathematical model of "what their needs and wants are" (Meta, 2018). The random act of getting to know people is therefore supplanted by an active influence. In this sense AIs are social actors, but do not appear within the social structure as we would normally understand it. They do this by defining what we see online and whom we meet on social platforms without explicitly searching for them (Lazer, 2015; Vosoughi et al., 2018). While each of these examples can be theoretically bypassed by the user, thus indicating a weaker framing effect than language as a symbolic order, they represent the immediate content of our online life — the 'how' of social reality. A new type of social actor arrives on the scene so to say. However, this actor is very untypical, as it is solely located in the social unconscious or the symbolic order as Mathew Flisfelder (2012: 64) points out. Its social action of curating content is located in the symbolic order, which it organizes without adding something-to-be-organized to it. The common idea of language as communication only considers explicit language as content and meaning. Language encompasses more than this content level, as Martin Heidegger has already highlighted in Being and Time, where he considers language to be the "articulation" of the structure and frame of what is "thematic" in it (Heidegger, 1996: 140). We find comparable ideas in a broad range of modern thinkers (e.g., Wittgenstein, Habermas, Lacan) but can also show empirically that a specific language frames the same problems in different ways, thereby leading to different conclusions (Winskel et al., 2016). Such framing can be discussed on several levels, some more strongly oriented on specific thematic ideas (like biases), others more structured by the ontology of a field. In this new social field that is generated when people are connected by positively identified 'interests,' existential and structural connections might play a role, yet they are not central to establishing the link between social actors. In the end, these curated links have an explicit primary purpose, namely, keeping the user online for as long as possible and thus maximizing both their exposure to advertisements and subsequent click-throughs. Central to these content feeds is therefore a phenomenon of binding the user to their interests and making sure they are connected to other people who either engage with the same content or generate it themselves. These feedback loops are fundamental to the social field that emerges here (Ge et al., 2020: 2,269–2,270).

An often-discussed effect of this supplantation, which describes the substitution of existential experiences through media (a concept introduced by Salomon, 1994), is the spread of 'echo chambers' in which individuals with comparable social profiles are grouped together and reinforce their social interests through contact with other individuals of comparable interests, thus creating self-reinforcing social communities (Karlsen et al., 2017; Zhu et al., 2021). These echo chambers are already proving to have a tangible social impact (Levy and Razin, 2019), and different feed algorithms can be shown to produce different types of echo chambers (Cinelli et al., 2021). Such variations suggest that this dynamic is not simply a question of epistemic isolation by dismissing counterevidence to our beliefs (Fantl, 2021; Santos, 2021); rather, it is actively influenced by the supplantation mechanism of algorithms. Which means that algorithms are increasingly responsible for defining the frame of reference through which we approach our social reality.

It might still seem sensible to argue that the effect of echo chambers is overstated and that it is not difficult to escape them (Dubois and Blank, 2018). However, if we take echo chambers as a symptomatic aspect of our modern reflexive identity (i.e., an extreme byproduct of the same structuring element that makes up 'normal' behavior), we can assume that the last decade, where algorithms more and more intrude into the social as interlinking agents of social connections, only constitute a starting point. Why to consider mostly the last decades? Because what Michael Wheeler calls "online intelligence", i.e., the flexible and fluid real-time computerized judgement of data (Wheeler, 2005: 13) is a rather young phenomenon, especially in its effect on social structures. The idea of a "starting point" is supported by the increasing influence that social media has on constituting younger identities (as indicated by Raiziene et al., 2021). But in the best tradition of psychoanalytic research, we can nevertheless approach these symptoms by discerning the structural level that enables them. This means it is not the echo chamber, which is in focus, but the socio-logical structure that produces identities solely through mathematical models of "wants and needs".

However, this new social logic is not entirely consistent, even if it might appear as such. As software, it malfunctions on an ontological level (Floridi et al., 2015), thereby producing 'bugs' and mishaps. This is especially relevant as there are limits to machine learning's own applicability as the Salganik et al. study indicated. It means that while AIs do emerge as social actors in the sense of structural actor, they offer no help in discerning their own effects. A malfunction of AI in social terms will be hard to differentiate from its intended behavior. This question of whether we should interpret a certain behavior as a malfunction, or a deliberate function of AI can also not simply be answered by reverse engineering the software itself. First of all, this often will not be possible because the owner of the software would regard such an act as a violation of trade secrets. Second, the algorithmic is a fundamental mode of organization, originally based in repetition and the written word, as machine learning repeats, writes, and organizes knowledge and data

without any subjective insight as to how or why (Durán and Jongsma, 2021; London, 2019). This 'black box' of AI—that is to say, the inability to discern its inner workings—is a fundamental structure of modern AIs and a problem even for the discipline in designing these machines themselves (Arrieta et al., 2020). However, manifest inconsistencies of the structural level of language—the symbolic order in psychoanalytic terms (Žižek, 2012: 2)—are not a hindrance to AI as social actor. Quite the opposite, in fact. Since we cannot fully discern this problem by approaching the technical aspect alone, there is need to further discern the difficulties this socio-technical interaction produces.

This leads to a core problem of this theory of technological supplantation: There is so far very little theoretical research on this subject and its framing, especially as the most active forms of such supplantations have only been in effect for a very short period as feed algorithms and machine-based curation of social links is a new phenomenon. There is a pressing need to discern how to act in this new social field but very little indication as to how to do it, since we know little about its structures. Probing an emerging field and offering an initial framing is a genuinely philosophical task, and this paper will therefore try to offer a logical dissemination of the social interaction of AIs. This probing will orient itself using Lacanian logic. This logic-centered approach might seem overly formal but given that algorithms cannot be understood without their genesis in modern logic and mathematics, it is not avoidable. This also means that there will be a focus on the structural foundations of algorithms and machine learning, which are used to create social interactions online, not primarily on their effects. In this sense the paper intents to open up a new way to use Lacanian psychoanalysis with regard to AI. The works of Isabell Millar (2021), André Nusselder (2006), Jacob Johansson (2018) of course contribute to the AI discussion, but we would argue that these approaches operate on a vector that starts with the fantasy of AI and applies its Lacanian discussion of it to algorithmic intelligence. This is important but encompasses only one side of the Janushead of AI: the phallic appearance of AI, as Clint Burnham (2022) called it. The other approach would be to start with the logico-algorithmic side and its material structure of calculation and discuss it on the basis of a continental understanding of Logic (based centrally in Lacan and Freud, but also Heidegger and Badiou) and approaching fantasy as the end-product of this. Millar touches upon this in her discussion of Omega Numbers (Millar, 2021: 23–27), but there is no in-depth discussion or follow-up of this problem. Nusselder also skirts this problem in "Interface Fantasy," but considers it to be a question of interfaces as a fantasy (Nusselder, 2006: 63). Matthew Flisfeder (2021) approaches this problem directly but doesn't discuss the algorithmic Big Other's inherent inability to produce what its phallic image might imply. Johansson indicates the perverse nature of big data (Johanssen, 2018: 141–167), but doesn't bind it to the actual essence of computing: modern logic and applied mathematics. In this sense, the works of Millar and Nussefelder did not provide a foundation for the presented analysis but are those of another house. Johanssen and Flisfeder point in the same

direction in their theoretical reflections, but essentially accept the claim of rationality by the Anglo-Saxon tradition of logic: that even the computer's logic is if not identical with the symbolic as discerned by Lacan it at least doesn't contradict it. However, the analysis that a fundamental "misrecognition" happens in big data, which Rambatan and Johanssen (2022) made, formulates the direction of the presented paper. Unfortunately, Rambatan and Johanssen to not work through the symbolic structure of this misrecognition. More general Lacanian works on online culture, which encompass a broader sense of aestheticized or technologically mediated communication, also do not offer an insight into the logico-algorithmic base of this new social field.

We also do not wish to join the discourse on the point at which an AI becomes truly intelligent, nor do we wish to discuss strong AI claims, as of now, this is a fantasy and should be discussed as such. Instead, let us look at the social situations that algorithms create on an abstract basis of the logic we can discern there. This means we are utilizing Lacanian psychoanalysis neither as a form of social analysis, nor in a clinical sense, but as "the method that proceeds with the deciphering of signifiers without concern for any form of presumed existence of the signified" (Lacan, 2006: 630), i.e., as a logic of the unconscious. This logic of the unconscious operates on a central insight that psychoanalysis approaches practically under the concept of castration: there is no symbolic unity. Consequently, every discourse or use of language is oriented on a gap or void. In this sense, it operates on the exact opposite of schools of thought like system theory or positivism: "I do not accept […] that every science should refer to a unitary, or world, system" (Lacan, 1998: 8). This radical insight, which is onto-logical, is what needs to be applied to the AI discourse. In very basic terms, it means that we must break with the truism that "the basic architecture of the internet is one based simply upon sharing information" (Flisfeder, 2021: 54). Information theory is not free of ontology, which is simply accepted if we assume that a "Symbolic order" is implemented by algorithms (Flisfeder, 2021: 104). First, we should be aware how the algorithm of computer sciences is not fully within the symbolic as Lacan designates it and what that means for social interactions curated by computers.

Ontology should then be used in the precise sense that Heidegger gave. It is not a discussion of a specific ontic problem, i.e., one that already presupposes a specific ontology, but a discussion of the ontological dimension itself. Lacan has already indicated that the computer will have difficulties approaching the symbolic as such, despite being structured by the symbolic:

> "[…]it is not because it lacks the supposed virtue of human consciousness that we refuse to call the machine to which we would attribute such fabulous performances a "thinking machine," but simply because it would think no more than the ordinary man does, without that making [it] any less prey to the summonses [appels] of the signifier." (Lacan, 2006: 45 [translation corrected by the author])

What is the problem on a purely formal level? It is a use of "symbolic law as the purely positive production, rather than repression, of reality and its desires" (Copjec, 1994: 23–24) mirroring a critique that Joan Copjec mounted against Foucault. Without acknowledging this critique, we might assume that the "new algorithmic identity" (Cheney-Lippold, 2011) through cybernetics is actually a better representation of the real. But computers in general operate within a logical frame that assumes negation not as virtual, but only as a specific negation, in difference to the negations that Lacan called frustration and castration, which introduce logical indeterminate fields. This needs to be detailed further.

3   SYMPTOMS OF THE 21ST CENTURY

Both the genesis of AI out of modern applied mathematics and the structuring of communication by AIs indicate a problem that in this form has mainly been discussed in psychoanalysis—that of the social dimension of the unconscious. I wish to take up a specific concept to discuss this further: the *algorithmic unconscious* as introduced by Luca M. Possati (2020). This algorithmic unconscious indicates that the *repressed* that forms and structures an algorithm is of vital importance to its social effect. This does not mean that one should attribute any consciousness to it. The situation is much to the contrary, since the unconscious is not 'inside' our head in Lacanian terms but rather intersubjectively exists 'outside' of us, so to speak, as the implicit structure of language and the forms and frames of our habitus. Much more importantly however, it includes not only this explicit dimension, but also the virtual excess that is introduced by the symbolic. With psychoanalysis as our guide, we can approach not only the behavior of non-human social actors (Rahwan et al., 2019) but more importantly highlight a distinct techno-social interaction within the implicit organization of social realities. The main structural difference to the human social actor that quickly becomes evident at this stage is that the algorithmic is purely unconscious and acts as a social actor primarily within this unconscious structure of the social.

While this might initially sound strange, we must carefully parse what this actually means. To begin with, psychoanalysis is not used in a therapeutic sense here but more as a theoretical framework that will allow us to *discern and distinguish* certain empirically describable social elements of algorithms. What are these elements? Possati offers us three applicable repressed dimensions: First, there is the mathematical formula as an "opinion embedded in math" (O'Neil, 2017: 50). This means that the programmer of the algorithm *embedded their own perspective in the program* and model that has been programmed. This is the *most superficial level of its repression* and still somewhat easy to grasp. In theory one could easily conduct empirical research on the opinions of such programmers. Such studies would have to be specialized towards certain programs and mathematical models, of course, but with the sociology of science there is a discipline that can approach this problem directly. Most interesting here would be the implicit framing of such

opinions to approach the unconscious that structured the social dimension of the programmers' models. Ideally, such studies should focus on major content curators like Google and Facebook, but it seems unlikely that these firms would agree to let scientists discern their implicit structures. It also raises the question of whether or not they would try to influence the results. At any rate, since neither the AI itself nor its public presence has access to the perspective that it assumes, this dimension of its workings is repressed.

A *second and much more important level of repression* is found in the *logical structure of the algorithm* itself. Modern predicative logic has not been brought to the world in a moment of divine epiphany; rather it is the result of a scientific discourse that has its own traditions and conflict—most notably among the logicians themselves (Priestley, 2011: 125). Modern computer science is also a specific offshoot of this tradition that has to operate within the physically possible, which means that certain elements of the ontology of mathematics (e.g., absences) can only be considered within the limits of physical computation. Negations are most important here, as two (frustration, castration) of the three negations that Lacan introduces (Lacan, 2020: 51) can only be thought within a purely symbolic space that is not based on a systematic unity but can only produce unity as a somewhat precarious unary trait or "count-as-one" as Badiou calls it (Badiou, 2006). However, digital computing only operates on negation as either a failure of a specific (positive) model, for example in evolutionary algorithms (Sloss and Gustafson, 2020: 313–320), i.e., the "fitness" of a count-as-one in relation to its *Umwelt*, or privation as 'missing data' (cf., e.g., Chai et al., 2020). All these examples are variations of privation. Hence, the assumption that it is only a lack of complete data, which hinders social analysis.

However, as a formal science neither mathematics as such nor logic is bound by the physical, instead it operates within the purely symbolic, only limited by the internal consistency or necessity of its arguments. This is particularly evident when we consider certain strands of reasoning that have been excluded in the tradition of logic, namely, that of Heidegger, Lacan, and other continental philosophers who argued that nothing can be determined without accounting for these voids/absences in its structure. This tradition has had a minor role in the AI research conducted by Hubert Dreyfus, who disrupted and improved the AI discourse through a Heideggerian critique (Dreyfus, 1979). Heidegger, who studied mathematics, physics, and philosophy, is the most prominent critic of the traditional logic and focused much of his oeuvre on criticizing Aristotelian concepts of logic (see, e.g., Heidegger, 1976a, 1984, 1998), which are still the basis for much of modern predicative logic.[1] Nevertheless, there are other important authors we must consider here too. The Lacanian reinterpretation of the Aristotelian square, for example, has

---

[1] It should be noted that Aristotle did not allow for any kind of true absence. Even the concept of zero would have been impossible in his concept of logic and mathematics. Compare Kaplan (1999) and Rotman (1987).

not been widely discussed outside of continental philosophy (Grigg, 2005). In addition, the idea of the empty set in the sense of a foundational void as posited by Alain Badiou in his seminal work *Being and Event* is most likely not reproducible in technical systems (Badiou, 2006: 187–190). This has two implications. First, since there is neither a universal standpoint nor a metascience, the way computer science operationalizes mathematics is within the same problem of application that other sciences have. Computer scientists need to discern why and how their specific mathematization happens and where the differences between formal mathematics and the specific application of mathematics lie. Second, because the philosophical discourse on logic was essentially split for most of the last century, computer science as such has mostly cleaved to one tradition. This marks a second level of unconscious and repressed structure that organizes AI today: Its theoretical frame of reference is not universal but based on a specific perspective and tradition. *More importantly, it is based on a repression of the central logical relation that is needed to understand social structure like discourses, if we follow Lacan or Badiou.* The scientific reconstruction of this second unconscious level should be done through a philosophical approach to the history of ideas relevant to the specific development of modern technical algorithms. Again, we can approach this unconscious level according to a preexisting—yet in this regard largely unutilized—tradition of thinking. Unfortunately, a philosophy of computer science that is not deeply entwined with the ontology of the Anglo-Saxon tradition of logic does not really exist right now, so the prospects of a thorough analysis are low.

The *third level* and perhaps the most difficult to approach, is the classical psychoanalytic problem of the *reflexive structure that algorithms create if they are used as curators of social links* and through which we understand ourselves. This is strongly entwined with the second level of repression, because psychoanalysis as a logic of the unconscious makes explicit what is excluded in the second level of repression. This is the social in terms of the Lacanian big Other. However, since the big Other is more than what AI curation can influence, we need to specify the influence computation has on the creation of the symbolic order. This level partly includes the other two, as the sociogenesis of modern AI partly constitutes the specific type of operation that happens on this level. Whereas the other two structural levels might have made it appear as if AI should primarily be treated like a manifestation of the social, on this level we need to take AI seriously as a social actor. Here we need to approach the algorithm as a symbolic mirror that constitutes our identity. This means that we have to discern the logic that is at work here.

4   THE MIRROR OPERATION OF SOCIAL IDENTITY

Possati also notes that the algorithmic in social networks and other digital communication is structured as a part of the Lacanian mirror stage, and I agree with that notion. However, the mirror Possati references is the imaginary mirror of early childhood, not the symbolic mirror that later distinguishes and structures different

forms of subjectivity. Thus, we need to leave behind the imaginary mirror that Possati discusses. What interests us instead is the symbolic structure of the mirror, for which Lacan offers the following formula in Seminar X (for a more detailed analysis of this formula, see Heimann (2022)):

$$\frac{A}{a \mid \$}$$

This is the so-called pervert's formula, which marks a simplified baseline subjectivity or normality. We can understand this as a post-transcendental approach to the subject-object relation in which the real biophysical body identifies itself with a system of signifiers. The relation of these absolutely distinct elements is constituted by a social system. Two things should be noted in advance: The body is absolutely exterior to the signifier, it is not as such intelligible, but made intelligible by signifiers. Secondly, the relation between the signifiers and the body is only possible via a mediator (the mirror). In this sense the optical mirror relation is transposed here into formal terms. There are more complex formulae to approach this problem, but the pervert's formula is helpful to demarcate it. We should use this mirror formula here as a basic structure of reflection to understand what happens if a computer-based logic intersects with this reflection. Let us briefly distinguish the variables used here.

(a) marks the original, the human object that constitutes its own identity through the mirror. However, the (a) also stands for the *objet petit a* (object small a) in Lacanian terms, which is often called the material remainder. Hence (a) should be understood as both. In a Non-Lacanian sense, it can be read as a formalization of the Kantian thing-in-itself; as an object not expressible in language and only expressible in a formal calculus as a symbol for the void (Lacan, 2014: 39). To demarcate why, one needs to consider the distinction that Lacan makes between signifier and signified. The signifier, the formal structure of language, never relates to its signified but only to other signifiers. This basic structure of language, translated into common concepts, simply means that the word one uses to signify something has no inherent relation to this thing. Instead, common language is essentially structured by certain practical language games, to use a widely known concept in Wittgenstein's later philosophy. Therefore, the individual human has no identifying signifier, and the (a) therefore marks here an indeterminate and meaningless symbolic excess (not the sublime), or in formal terms an absence that is not intelligible as such. Because of its imaginary inexpressibility, it is something that can only appear as a counterfactual symbolic element when we push language to its final end. It's important to note that this indeterminate is exactly what pushes Lacanian logic in a radically different direction than the analytic philosopher's approach. By centering his logic on the *objet petit a* it becomes impossible to take on a systemic approach. Instead, the unity or system of a symbolic structure is always structured around this remainder. Quite notably, this element of excess is necessary if we wish to formalize the endless process that is desire and its inherent difficulty

to stay put with a specific object. For a more detailed discussion of this *object petit a* and the ontological difficulties it introduces, see Zupančič (2017: 74–139)

($) is the mirror image in which this impossible-to-know-object actively tries to know itself. This prospect is bound to fail from the start, but it fails in a productive way, so to speak. It fails because ($) is a signifier and therefore on a completely different ontological level than what it tried to understand, which is indicated by (a). That is why ($) is also called the "subject barred by the signifier." And yet within this failure there is also an aspect of success, because the relation between (a) and ($) now allows for a basic structure of identity, however this relation requires mediation through (A). This means that ($) is also a 'visible' failure, as any mirror image requires the mirror operation for identification.[2] Without going further into the Lacanian theory of the subject that does formulate more complex formulae of (neurotic) subjectivity, we will further discern the space (A) that interests us most, as it is where the social action of AI would intervene.

(A) with the vertical line marks the mirror itself. While the mirror (|) is formally structured as an operator, the (A) indicates a specific influence, which structures *how* the mirror inverts the (a) into ($). The 'mirror image' ($), comparable to the function of a visual mirror, is a virtual object. It is the social identity that we constitute through language. This means there is no material basis for this object. It exists as a symbolic structure enabled by the socio-linguistic mirror operation (A). The (A) usually is used in Lacanian theory for the big Other, the complex structure of the intersubjective unconscious that organizes how we use language. This big Other has several elements that are partly derived from its original genesis out of the infant's situation. It structurally acts as a guarantor of the identification, given that the factual non-identity of the mirror images (as is the case with the optical mirror) is superseded by the mirror operation.

Lacan developed this originally from the infant's experience with a mirror, as Possati describes, but the mirror operator functions on the symbolic structure of language. This mirror itself, as Lacan proposed, always acted on the basis of algorithmic organization in the position of A, but in Lacan's time these algorithms were assumed by him to be buried in the implicit and tacit knowledge of whole societies and as solely based in the mathematical core of language. He assumed that the organizing structure of algorithmic forms can be found in the unconscious as a linguistic element that organizes knowledge and personal identity in a fundamental sense (Lacan, 2006: 21–48). This means that the algorithmic unconscious is older than its current appearance as a technical artifact. However, today another algorithmic unconscious appeared that operates much faster than the algorithmic social unconscious and it operates differently because of its foundation in machine logic.

---

[2] To grasp the inherent problem of a mirror operation constituting an identity, look at your own hands to realize that the only way these can be identified with each other is through an intermediate mirror operation, an inversion that changes their appearance. The big Other (A) as the symbolic order is therefore that which constitutes the identity of otherwise different enantiomorphs.

An algorithm that curates and enables a certain social field for me acts within this space formerly organized by the implicit algorithms of the big Other that Lacan describes. It does so, without ever needing to breach the explicit level of language because it chooses and curates, based upon its own (structurally lacking) ability to discern my desires, what I see. Therefore, it frames and organizes the explicit content as noted by Flisfeder (2021: 64). This is what the (A) in the depicted formula indicates, the specific structure and organization that enables me to approach a culturally and ideologically framed mirror image ($).

5   THE ALGORITHM AS MIRROR

If we assume that social algorithms function as a mirror for the identification process of normality, then we have to confront a basic problem of representation. For a start, a *meaningful* order can arise out of totally random data. This is a basic idea that Lacan introduced early in his seminars and it highlights the problem that the signifier is not bound by whatever it intends to signify. It is a common concept in the philosophy of language and can be articulated by Wittgenstein's concept of language games as well as Heidegger's 'the they.' This means that something can appear to us as totally rational, solely through the symbolic order introduced by language, despite not possessing any internal order.

For Heidegger, this basic structure of practice-bound linguistic structures is limited by the positive function of empirical and existential failure (Heidegger, 1967: 242). As these language games can fail, we still have a possible means of accessing reality as such. However, in Heidegger's philosophy there is no distinct idea of resistance (*Widerstand*), which is the second element to be considered here. In Lacanian psychoanalysis resistance is oriented on the disavowal of (a) as a remainder of this identification (Lacan, 1993: 242) or as an impasse of being (Zupančič, 2017: 22). Identities that are solely based in the pervert's formula or the adherence to the intersubjective mirror would be negated if the unintelligibility of (a) is made explicit, because the failure that constitutes this identity would be laid bare. The consequence of this is that in the psychoanalytic theory of subjectivity, this subject is essentially a detour to avoid this. This idea was first introduced by Freud in his seminal text "Beyond the Pleasure Principle" (Freud, 2010; Zupančič, 2017: 96–101). The goal of the psychoanalytic cure is therefore to integrate this remainder as such—not as a negation but by assuming the logical non-all as the subject position (Dulsster, 2022: 15; Žižek, 2012: 745–750). Subjectivity is then an awareness of the lack that the signifier introduces but without the (pathological) desire to get rid of it. However, as Lacan already remarked, this is not true for the machine:

> With a machine, whatever doesn't come on time simply falls by the wayside and makes no claims on anything. This is not true for man, the scansion is alive, the ego in Freud's theory and in the technique of psychoanalysis and whatever doesn't

come on time remains in suspense. That is what is involved in repression. (Lacan, 1991: 308–309)

What does this mean in terms of the mirror operation? To begin with, it allows us to explicate an argument that there are certain elements of rationality that computers are unable to reproduce, which Slavoj Žižek (2021) only hinted at. We need to ask what the difference is between a supplanted social structure and a classical social structure in formal terms. An existentially founded community is at risk of the pitfalls of the detouring subject in that it introduces strife, discordant ideas and material failure, which all create suffering, but this also enables us to accept our castration, which is the classical name for the effect of (a). That is why the neurotics formula of this reflection includes the *objet petit a* as part of the mirror image:

$$A$$
$$S \mid a \, \$$$

The neurotic includes the lack as part of his mirror image and while even this inclusion still can be repressed, it nonetheless offers a path towards its integration. While the pervert is the subject identity structured by the pervert's formula, which excludes the remainder (a) in his symbolic identity. The pervert's identity is still based on it, as they assume the position of this object of desire (Fink, 1999: 128). In more general terms, this existential situation forces us to constantly redefine our relation to the material reality because we fail to grasp it, as the remainder (a) marks our symbolic order as incomplete and inconsistent. The main difference for a supplanted community, i.e., one that is constituted by a logic that cannot operate with a representation of (a), would therefore seem to be that resistance and repression are strengthened to a point where the structure of castration (i.e., the failure/contingency of ($) and the problem that (a) marks) is no longer part of the identity structured by it. This means that a social identity is no longer created as a reflection of the existential basis (a) but instead created purely through the symbolic structure of (A). To propose a metaphor, the mirror operation vanishes or becomes a display that creates a virtual image solely out of already existing symbolic structures while the structural necessity to redefine those elements is much less pressing.

How can one assume that the AI-supplanted social structures would strengthen resistance and repression? We can assume that this displayed image strengthens both because the structure of predicative logic and applied mathematics upon which AI is built cannot access the formal problem that is bound to appear with (a). This is very well documented in Carnap's critique of Heidegger and its regular reoccurrence. Carnap criticizes Heidegger for using the concept of negation as an active virtual possibility ("das Nichts nichtet") instead of using negation only as the negation of existence (Carnap, 1959). This is in Carnaps view a misuse of the negation as a determined negation of something. He insists that the only logical approach to negation is a privation, that is, the specific negation of an existential

judgment, disallowing psychoanalytic concepts of negativity as well. Notably, this critique is a repeating phenomenon. It appears with some regularity: now and then someone attempts to demonstrate that Heidegger's work on negativity isn't real philosophy but essentially bad poetic scribblings. This has been well documented by Stephan Käufer (Compare Käufer, 2005: 146). Modern computer science is a specific offshoot of this tradition (Priestley, 2011). Accordingly, the active avoidance of a more complex concept of negation is still active within modern computer science. Therefore, identities are grasped by AI as a set of positive identifiers (cf., e.g., Y. Wu et al., 2017), whereas the excluded (a) that is central for subjectivity cannot be expressed. Alenka Zupančič notably used this joke to explain it:

> A guy goes into a restaurant and says to the waiter, "Coffee without cream, please."
>
> The waiter replies, "I am sorry, sir, but we are out of cream. Could it be without milk?" (Zupančič, 2012)

In terms of our identity, the "without" (or lack) that is used in this joke as a positive element is the unintelligible that is our concrete, individual, and real body that doesn't fit the symbolic. The remainder (a) in this case is the real person as a thing-in-itself being excluded from the structure of linguistic representation by not being a signifier but rather an absolutely necessary element of the (necessarily failed) identity. This means that in psychoanalytic terms a personal identity is always a combination of certain positive identifiers and the specific way in which these fail to grasp the actual individual. The "without" is therefore a positive element of its own. However, this "without" only appears on the basis of this failure, which is why the identity still succeeds in part.

The reason for the failure of algorithms to reflect this lies not simply in formalization or mathematization as such. One does rely on Lacanian formulae to explicate it. However, in the specific intellectual tradition that gave birth to modern logic and AI, this "without" that is marked is considered irrelevant. The concept of zero that is applicable to material computation is either the neutral position or the absence of a change, neither of which is identical to the absence of (a) as the inconsistency of the symbolic identity. In stark contrast to this, Heidegger, Lacan, and the continental philosophical tradition that followed them did not exclude the indeterminate from logic, and their formalizations and mathematizations of the social reflect that. In Badiou's words, "None of this [the distinction between formal and empirical sciences in the Anglo-Saxon tradition of logic] was consistent with the clear Lacanian doctrine according to which the real is the impasse of formalization" (Badiou, 2006: 5). This impasse is the (a) as a part of the psychoanalytic mathematization of the unconscious and the inclusion of the real into formalizations of the subject. This Lacanian approach is preceded by Freud's focus on the speech uttered by the analysand as the empirical basis for

psychoanalysis (Hainzovich, 2002). Freud noted early on that the structure of the unconscious cannot be discerned with a classical approach to logic (Freud, 1942: 317–319).

This, however, does not mean that there is no formal structure to this "without", thus an empirical "logic of the unconscious" must be considered from a basis other than the predicative sentence of the Boole/Frege/Russel line of thought. Lacan's answer to this is the mirror operation as a basis for the identity presupposed in classical logic (Heimann 2022). This is the reason that the *objet petit a* (a) that Lacan introduced, as a structurally and ontologically unintelligible and indeterminate element, is logically excluded from any calculation that is based on constructed and determinable sets. However, the very idea of determination and identity and thus counting relies upon an indeterminate (*the* nothing) as Heidegger first showed (Heidegger, 1999: 82–96). We also can extract this from Frege's *Foundations of Arithmetic*, as demonstrated by Jaques Alain Miller (1977), despite Frege being one of the founding authors of the modern logic that excludes this indeterminate. Expressible in the idea that "only the measurable is real," the now classical approach to logic excludes its own metric and axioms (i.e., the conditions of its own consistency marked by this absence). This leads to the very concise problem that a mathematization of the social, despite being a clear goal of Lacanian psychoanalysis, cannot be simply transferred from the existing mathematization of nature. Instead, the indeterminate and unintelligible as a central element of social structures and personal identities needs to be taken seriously if we wish to mathematize it.

One can therefore assume that this difference in logic, this lack of the access to the unconscious, acts itself as an unconscious structure of the social action that AIs produce. This might sound highly paradox: to claim that (1) the machine has no access to the unconscious, but at the same time that (2) social AI's act solely within the unconscious and determined by the repressed traditions of logic, which acts as a material unconscious (3). This lack of lack is an explicable problem that Lacan already indicated in another aspect regarding the capitalist's discourse as it disavows the split in the subject (Vanheule, 2016: 7). It means that (1) modern digital computing has no way to include the unconscious into its formalizations and thus can only create reflections of our action which actively exclude the unconscious. However, the way these reflections are created – by curation of content – is (2) purely located on the unconscious level of our reflections as the framing of reality by a severely stunted symbolic order. For the machine, the repressed is simply excluded, it cannot enter the computation at all. This doesn't mean however, that it doesn't impact the machine's actions in the social space, where it is no longer singular, but effects other social actors. By constructing the mirror image ($) without any link to castration, its production is then determined by this exclusion (3). This then is what we call the social AI's supplantation, the construction of a socially mediated identity ($) without any possible relation to (a).

What we then find in the AI-supplanted identity is the 'obscene immortality' that Žižek notes in our consumption of video games and how subjectivity is displayed in modern science fiction movies (Žižek, 2017). This obscene immortality is the exclusion of the structural element of castration. And, of course, this has effects on the subject constituted by it. The subject appears to itself as immortal in the sense that it has no concept of its symbolic finitude or that its identity is essentially a false one and produced via a failure. We cannot assume that this is relevant to every subject, but it might offer a basis for understanding the symptomatic identity structures that we see in the echo-chamber social identity. This supplantation might also increase the vulnerability of ideological framings as it removes the inherent lack that is apparent in the mirror operation. The algorithmically curated identity thus produced without an indeterminate remainder appears more complete than the regular identity. What is interesting here is that the opposite of the Freudian problem appears. While castration is for Freud a central element of the subject's identity that it has to face and acknowledge, the identity that AI as social actors produces is structurally uncastrated because the repressed remainder is impossible within the logic of the machine.

6   CONCLUSION

This turns social AI not only into a technological manifestation of social structures but into an object that is equiprimordial in the realms of technology and the social, thus calling for an approach that is focused neither on the technical nor the social but a theory of a material unconscious. The reason for this is that the material structure of computation changes the symbolic order it can produce. This is because what we see in AI as a social actor is a complex interplay between technical and social structures. By identifying the inability to formalize the lack in social AIs and their innate inability to reflect anything but an uncastrated mirror image, we can also identify a necessary field of action for further research. This creates a situation that is (at least formally) comparable to the infantile inability to comprehend the void that (a) creates (Stavchansky, 2018: 10). However, the child still symbolizes this lack unconsciously, so the identity that is partly AI-structured introduces a new mirror relation that removes this determined negation altogether.

This also means that another field of social action has opened up with the automated sociogenesis of milieus. Echo chambers, for example, are extremely hard to maintain in offline communities because they require political suppression and active policing of adverse opinions, yet they are a new normal in online communities. At the same time, this is not an isolated phenomenon somewhere in the virtual space. Online communities increasingly affect offline communities, as is visible in the widespread conspiracy theories that surfaced during the COVID-19 pandemic (Allington et al., 2021; Earnshaw et al., 2020) or in recent democratic elections (Faris et al., 2017). They are not isolated phenomena of a virtual space but rather active elements of the social space as a whole, including offline and online

communities. This also means that it is also not solely a problematic or pathological field. However, the pathological phenomena of this new field, such as echo chambers and the new variants of symptomatic behavior, do provide a rich area of research. The analysis detailed here aims to provide another viewpoint on this, by taking seriously the foundational logic discourse which constitutes the logic of modern computers. The difference between Lacanian or continental philosophy logic and the logic as utilized in computer science is, however, not one to be found in the complexity of models, but in the simplicity of foundational decisions. By constituting distinction and identity on the basis of an indeterminate, as Heidegger first proposed in "What is Metaphysics?", the foundations of computer science as it is physically determined might be already too complex to account for the indeterminate void found in the symbolic order of social life that Lacan describes.

Still, the possibility of creating social bonds independent of cultural upbringing can in turn allow for new types of social structures that might not only show pathological dimensions but also allow for new types of social action altogether. People who have been atomized by the structural elements of their economic or political situation can connect, and we see such connections being central to the creation of, say, unions of platform-based employment situations (Katsabian, 2021). What we are witnessing here is thus the rise of a new social field, one that operates according to rules that are very different from those we know from the classical mirror relation.

FUNDING STATEMENT AND ACKNOWLEDGMENTS

REFERENCES

Allington D, Duffy B, Wessely S, et al. (2021) Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. Psychological medicine 51(10): 1763–1769.

Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58: 82–115.

Badiou A (2006) Being and event. London: Continuum.

Burnham C (2022) Siri, what is psychoanalysis? Psychoanalysis, Culture & Society. DOI: 10.1057/s41282-022-00294-0.

Carnap, R. (1959) The elimination of metaphysics through logical analysis of language (trans. A. Pap). In A. Ayer (ed.), Logical Positivism. New York: The Free Press.

Chai X, Gu H, Li F, et al. (2020) Deep learning for irregularly and regularly missing data reconstruction. Scientific reports 10(1): 3302.

Cheney-Lippold J (2011) A New Algorithmic Identity. Theory, Culture & Society 28(6): 164–181.

Copjec J (1994) Read my desire: Lacan against the historicists. Cambridge, Mass. u. a.: MIT Press.

Cinelli M, Francisci Morales G de, Galeazzi A, et al. (2021) The echo chamber effect on social media. Proceedings of the National Academy of Sciences of the United States of America 118(9). DOI: 10.1073/pnas.2023301118

Dressel J and Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Science advances 4(1). DOI: 10.1126/sciadv.aao5580

Dreyfus HL (1979) What computers can't do: The limits of artificial intelligence. New York: Harper & Row.

Dubois E and Blank G (2018) The echo chamber is overstated: the moderating effect of political interest and diverse media. Information, Communication & Society 21(5): 729–745. DOI: 10.1080/1369118X.2018.1428656

Dulsster DGM (2022) The Reign of Speech: On Applied Lacanian Psychoanalysis. Cham: Springer International Publishing; Imprint Palgrave Macmillan.

Durán JM and Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. Journal of medical ethics. Epub ahead of print 18 March 2021. DOI: 10.1136/medethics-2020-106820.

Earnshaw VA, Eaton LA, Kalichman SC, et al. (2020) COVID-19 conspiracy beliefs, health behaviors, and policy support. Translational behavioral medicine 10(4): 850–856. DOI: 10.1093/tbm/ibaa090

Fantl J (2021) Fake News vs. Echo Chambers. Social Epistemology 35(6): 645–659. DOI: 10.1080/02691728.2021.1946201

Faris RM, Roberts H, Etling B, et al. (2017) Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election. Cambridge, Massachusetts: Berkman Klein Center for Internet & Society Research.

Fink B (1999) A clinical introduction to Lacanian psychoanalysis: Theory and technique. Cambridge, Mass.: Harvard Univ. Press.

Floridi L, Fresco N and Primiero G (2015) On malfunctioning software. Synthese 192(4): 1199–1220. DOI: 10.1007/s11229-014-0610-3

Flisfeder M (2021) Algorithmic desire: Toward a new structuralist theory of social media. Evanston, Illinois: Northwestern University Press.

Freud S (1942) Die Traumdeutung: Über den Traum. Frankfurt am Main: S. Fischer.

Freud S (1968) Gesammelte Werke: Werke aus den Jahren 1917- 1920. Berlin: S. Fischer.

Freud S (2010) Gesammelte Werke: Jenseits des Lustprinzips/ Massenpsychologie und Ich-Analyse/ Das Ich und das Es. Frankfurt am Main: S. Fischer.

Ge Y, Zhao S, Zhou H, et al. (2020) Understanding Echo Chambers in E-commerce Recommender Systems, Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval: 2261–2270. DOI: 10.1145/3397271.3401431

Grigg R (2005) Lacan and Badiou: Logic of the pas-tout. Filozofski vestnik XXVI(2): 53–65.

Harman G (2016) Immaterialism: Objects and social theory. Cambridge, Malden, MA: Polity.

Hainzovich S (2002) Freud's Pre-Analytical Writings and his Scientific Revolution. In: van Vijver G de (ed) The pre-psychoanalytic writings of Sigmund Freud: London: Karnac, pp. 207–214.

Heidegger M (1967) Sein und Zeit. Tübingen: Max Niemeyer Verlag.

Heidegger M (1976a) Logik: Die Frage nach der Wahrheit. Frankfurt am Main: Vittorio Klostermann.

Heidegger M (1984) Grundfragen der Philosophie: Ausgewählte »Probleme« der »Logik«. Frankfurt am Main: Vittorio Klostermann.

Heidegger M (1996) Being and time: A translation of Sein und Zeit. Albany, NY: State University of New York Press.

Heidegger M (1998) Logik: Als die Frage nach dem Wesen der Sprache. Frankfurt am Main: Vittorio Klostermann.

Heidegger M (1999) Pathmarks. Cambridge: Cambridge Univ. Press.

Heimann M (2022) The Mirror Operator: On Lacanian Logic. *The International Journal of Psychoanalysis* 103(5): 707–725. DOI: 10.1080/00207578.2022.2035732

Kaplan R (1999) The nothing that is: A natural history of zero. Oxford, New York: Oxford University Press.

Karlsen R, Steen-Johnsen K, Wollebæk D, et al. (2017) Echo chamber and trench warfare dynamics in online debates. European journal of communication 32(3): 257–273. DOI: 10.1177/0267323117695734

Katsabian T (2021) Collective Action in the Digital Reality: the Case of Platform-Based Workers. The Modern Law Review 84(5): 1005–1040. DOI: 10.1111/1468-2230.12635

Käufer S (2005) Logic. In: Dreyfus HL (ed) *A Companion to Heidegger:* Malden, Oxford, Carlton: Blackwell Publishing, pp. 141–155.

Johanssen J (2018) Psychoanalysis and Digital Culture: Audiences, Social Media, and Big Data. Milton: Routledge.

Lacan J (1993) The Psychoses. New York: W.W. Norton & Company.

Lacan J (1998) The four fundamental concepts of psychoanalysis. New York, NY: Norton.

Lacan J (2014) Anxiety. Cambridge, Malden, MA: Polity.

Lacan J (2006) Ecrits: The first complete edition in English. New York: Norton.

Lacan J (2020) The object relation. Cambridge, UK, Medford, MA, USA: Polity.

Lada A, Wang M and Yan T (2021) How does News Feed predict what you want to see? Personalized ranking with machine learning. Available at: https://tech.fb.com/engineering/2021/01/news-feed-ranking/ (accessed 20 September 2022).

Lazer D (2015) Social sciences. The rise of the social algorithm. Science (New York, N.Y.) 348(6239): 1090–1091.

Levy G, Razin R (2019) Echo Chambers and Their Effects on Economic and Political Outcomes. Annual Review of Economics 11(1): 303–328. DOI: 10.1146/annurev-economics-080218-030343

Littlefield AK, Cooke JT, Bagge CL, et al. (2021) Machine Learning to Classify Suicidal Thoughts and Behaviors: Implementation Within the Common Data Elements Used by the Military Suicide Research Consortium. Clinical Psychological Science 9(3): 467–481. DOI: 10.1177/2167702620961067

London AJ (2019) Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. The Hastings Center report 49(1): 15–21.

Mathiyazhagan S (2021) Field Practice, Emerging Technologies, and Human Rights: the Emergence of Tech Social Workers. Journal of human rights and social work: 1–8. DOI: 10.1007/s41134-021-00190-0

Meta (2018) How Users Help Shape Facebook. Available at: https://about.fb.com/news/2018/07/how-users-help-shape-facebook/ (accessed 20 September 2022)

Millar I (2021) The Psychoanalysis of Artificial Intelligence. Cham: Springer International Publishing; Imprint Palgrave Macmillan..

Miller J-A (1977) Suture, elements of the logic of the signifier. Screen 18(4): 24–34.

Nusselder AC (2006) Interface fantasy: A Lacanian Cyborg Ontology een Lacaniaanse Cyborg Ontologie = Interface fantasie. Zugl.: Rotterdam, Univ., Diss., 2006. Amsterdam: F&N Eigen Beheer.

O'Neil C (2017) Weapons of math destruction: How big data increases inequality and threatens democracy. New York: B/D/W/Y Broadway Books.

Pence HE (2019) Artificial Intelligence in Higher Education: New Wine in Old Wineskins? Journal of Educational Technology Systems 48(1): 5–13. DOI: 10.1177/0047239519865577

Possati LM (2020) Algorithmic unconscious: why psychoanalysis helps in understanding AI. Palgrave Communications 6(1). DOI: 10.1057/s41599-020-0445-0

Priestley M (2011) A Science of Operations. London: Springer London.

Rahwan I, Cebrian M, Obradovich N, et al. (2019) Machine behaviour. Nature 568(7753): 477–486. DOI: 10.1038/s41586-019-1138-y

Rambatan B and Johanssen J (2022) Event horizon: Sexuality, politics, online culture, and the limits of capitalism. Lanham: John Hunt Publishing Limited.

Raiziene S, Erentaite R, Pakalniskiene V, et al. (2021) Identity Formation Patterns and Online Activities in Adolescence. Identity: 1–16. DOI: 10.1080/15283488.2021.1960839

Rotman B (1987) Signifying nothing: The semiotics of zero. Basingstoke: Macmillan.

Salganik MJ, Lundberg I, Kindel AT, et al. (2020) Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences of the United States of America 117(15): 8398–8403. DOI: 10.1073/pnas.1915006117

Salomon G (1994) Interaction of Media, Cognition, and Learning: An Exploration of How Symbolic Forms Cultivate Mental Skills and Affect Knowledge Acquisition. London: Routledge.

Santos BRG (2021) Echo Chambers, Ignorance and Domination. Social Epistemology 35(2): 109–119. DOI: 10.1080/02691728.2020.1839590

Sloss AN, Gustafson S (2020) Chapter 16: 2019 Evolutionary Algorithms Review. In: Banzhaf W, Goodman E, Sheneman L, Trujillo L and Worzel B (eds) Genetic Programming Theory and Practice XVII: Cham: Springer International Publishing; Imprint Springer.

Stavchansky L (2018) Lacanian Psychoanalysis between the Child and the Other: Exploring the Cultures of Childhood. Boca Raton, FL: Routledge.

Vanheule S (2016) Capitalist Discourse, Subjectivity and Lacanian Psychoanalysis. Frontiers in psychology 7: 1948.

Vosoughi S, Roy D and Aral S (2018) The spread of true and false news online. Science (New York, N.Y.) 359(6380): 1146–1151. DOI: 10.1126/science.aap9559

Wheeler M (2005) Reconstructing the cognitive world: The next step. Cambridge, Mass.: MIT Press.

Winskel H, Ratitamkul T, Brambley V, et al. (2016) Decision-making and the framing effect in a foreign and native language. Journal of Cognitive Psychology 28(4): 427–436. DOI: 10.1080/20445911.2016.1139583

Y. Wu, Q. Lv, Y. Qiao, et al. (2017) Linking Virtual Identities across Service Domains: An Online Behavior Modeling Approach. International Conference on Intelligent Environments (IE): 122–129. DOI: 10.1109/IE.2017.29

Zhu J, Ni P, Tong G, et al. (2021) Influence Maximization Problem With Echo Chamber Effect in Social Network. IEEE Transactions on Computational Social Systems 8(5): 1163–1171. DOI: 10.1109/TCSS.2021.3073064

Žižek S (2012) Less than nothing, Hegel and the shadow of dialectical materialism. London, New York: Verso.

Žižek S (2017) The Obscene Immortality and its Discontents. International Journal of Zizek Studies 11(2).

Žižek S (2021) Blade Runner 2049: A View of Post-Human Capitalism. In: Neill C (ed) Lacanian Perspectives on Blade Runner 2049: Cham: Springer International Publishing; Imprint Palgrave Macmillan, 41-51.

Zupančič A (2012) Sexual Difference and Ontology. e-flux journal 32.

Zupančič A (2017) What is Sex? Cambridge, London: MIT Press.

NON-KNOWLEDGE IN MEDICAL PRACTICES: APPROACHING THE USES OF SOCIAL MEDIA IN HEALTHCARE FROM AN EPISTEMOLOGICAL PERSPECTIVE

Anna Sendra[a], Sinikka Torkkola[a] and Jaana Parviainen[a]

ABSTRACT

Social media has transformed how individuals handle their illnesses. While many patients increasingly use these online platforms to understand embodied information surrounding their conditions, healthcare professionals often frame these practices as negative and do not consider the expertise that patients generate through social media. Through a combination of insights from social epistemology and ignorance studies, this paper problematizes the distinctive understandings of social media between patients and healthcare professionals from a different perspective. A total of four ideas are introduced: (1) healthcare professionals see embodied knowledge that arises from patients' social media practices as uncomfortable knowledge; (2) healthcare professionals engage in several behaviours to preserve their authority and power in front of embodied knowledge created through these online platforms; (3) failing to consider embodied knowledge can have consequences not only in terms of trust between patients and healthcare professionals but also in connection with epistemological populism and the transition towards patient-centred care; and (4) media and digital health literacy could help healthcare professionals enhance the uses of social media in healthcare. Adopting this framework allows not only to offer valuable insights into how healthcare professionals manage patients' social media practices, but also opens new avenues to improve healthcare digitisation.

Keywords: social media; embodied knowledge; healthcare professionals; patients; ignorance studies; social epistemology.

[a] Tampere University, Finland

1   INTRODUCTION

In the realm of digital health, one of the most used technologies by patients is social media (Lupton 2018). These online platforms have the capability to provide informational and emotional support to those who experience illness (Sendra et al. 2020). Patients who used to have no answers now feel validated in the experiences of other people (Wagner et al. 2021). Despite their value, healthcare professionals[1] not only still use social media poorly (Hernandez et al. 2021), but also react negatively quite often when patients ask them about information found through these online platforms (Benetoli et al. 2018). A considerable amount of literature has established these distinctive understandings of social media between patients and healthcare professionals (e.g., Sendra and Farré 2017). Still, research to date has not yet determined how to improve the adoption of these online platforms in the work practices of the latter.

Starting from the premise that knowledge is a key feature in medicine (Martin et al. 2009), this theoretical paper examines how patients and healthcare professionals struggle with their unknowns and use their own knowledge resources differently in the age of digital health. We integrate ideas of previous studies on social epistemology (Fuller 2002) and ignorance studies (Gross and McGoey 2015; 2022) to provide new conceptual insights on how patients and healthcare professionals deal with these online platforms and the practices that accompany them. Although the research is theoretically oriented, we utilise previous empirical studies of all authors (Sendra and Farré 2020; Torkkola et al. 2019) about patients and healthcare professionals as users of social media in acquiring health information. Furthermore, we use empirical and theoretical findings from previous studies in the field of digital health (e.g., Lupton 2018; Sosnowy 2014; Ventola 2014) to highlight and deepen our conceptual framework. Especially, the study explores the illustrative cases of patients with long COVID and chronic pain to uncover how patients make their unknowns known.

1.1   Theoretical foundation

Historically, the relationship between patients and healthcare professionals has been based on superiority, where healthcare professionals are the individuals in possession of (expert) knowledge (Wagner et al. 2021). Although the paternalistic nature of their encounters took a turn towards patient autonomy in the 1980s (Killbride and Joffe 2018), the emergence of social media consolidated participatory healthcare (Sosnowy 2014). Along with this transformation, a type of expertise different from expert knowledge was introduced (Bellander and Landqvist 2020). As indicated elsewhere, the expertise that arises from patients' social media practices goes beyond evidence-based medicine (Sendra and Farré 2020). Consequently, healthcare professionals are increasingly guided "to new horizons of what is

---

[1] By healthcare professionals we mostly refer to nurses and general practitioners.

unknown" (Gross and McGoey 2015, p. 1). Epistemology encompasses this unknown under the concept of medical ignorance (hereafter, non-knowledge) (Whooley and Barker 2021).

As Whooley and Barker (2021) explained, non-knowledge can adopt different forms. To name a few, in medicine there can be *known unknowns* (unknowns that we are conscious of their existence); *unknown unknowns* (unknowns that we are not conscious of their existence); *tacit knowing* (unknowns that we do not know that we are conscious of their existence); *errors* (unknowns that we think we are conscious of their existence but that are erroneous); *taboos* (unknowns that we are not supposed to be conscious of their existence but that may be useful); and *denials* (unknowns that we refuse to be conscious of their existence) (Kerwin 1993). The bottom line of these categorisations is to highlight that knowledge in medicine is limited, where those things that we do not know are intertwined with those things that we do know (Kerwin 1993; Whooley and Barker 2021). As a result, "ignorance saturates all of medicine, from the biomedical laboratory to the design of medical technologies, from the clinical encounter to collective politics of health and illness" (Whooley and Barker 2021, p. 280).

While non-knowledge existed before the emergence of social media (Kerwin 1993), the use of these online platforms for health-related purposes has enhanced its visibility, particularly of known unknowns[2]. The problem remains that only few studies have examined how healthcare professionals deal with the non-knowledge that arises from patients' social media practices. In previous research, the relationship between knowing/not knowing has been explored under the framework of genetics (Wehling 2015), mental health (McPherson et al. 2020) or antibiotics (Will 2020). As for social media, scholars have focused on analysing how patients generate lay expertise through these online platforms (Bellander and Landqvist 2020; Maslen and Lupton 2019). Therefore, the aim of this paper is to illustrate how non-knowledge impacts the ways healthcare professionals are managing the unknowns that patients generate through social media. The discussion will be guided by the *structural conception of ignorance*. As indicated by El Kassar (2018, p. 302), this notion suggests that non-knowledge is administered in a way that "is not just rooted in the beliefs, epistemic vices and the outlook of the individual but also manifest in and maintained by social and institutional structures and mechanisms".

The following sections present four main ideas. First, we argue that healthcare professionals see the non-knowledge that emerges from patients' social media practices as uncomfortable knowledge (Rayner 2012). Second, we problematise why healthcare professionals fail to consider the non-knowledge created through these online platforms in their work practices (McGoey 2012, 2020; Stein 2020; Williams 2021a). Third, we illustrate that not considering the non-knowledge that emerges from patients' social media practices can have

---

[2] An example of known unknowns are not yet registered causes behind a disease (e.g., long COVID).

consequences for the organisation and delivery of healthcare at the micro, meso and macro levels (Akrich 2010; Nie et al. 2018; Numerato et al. 2019). Fourth, we suggest that media and digital health literacy could improve the uses of social media in healthcare for both patients and healthcare professionals (Jones et al. 2021; Marchal and Au 2020; Mather and Cummings 2019; Torkkola et al. 2019). The paper concludes with a summary of the contribution and the limitations of this study.

2 HEALTHCARE AND THE BIOMEDICAL PARADIGM

Healthcare professionals are arguably one of the most consolidated epistemic communities in society. Their expert knowledge has been acquired through long years of study and practice in healthcare organisations (Wagner et al. 2019; Wilkesmann 2016). This expertise is also constantly nurtured through multiple epistemic objects, such as blood tests or scans (Nerland and Hasu 2021). The result of these epistemic practices, which are rarely questioned, is patients' profound trust in healthcare professionals (Ahmed et al. 2020). However, the use of social media for health-related purposes is testing the traditional functioning of this epistemic community. Evidence-based knowledge production processes of healthcare professionals are increasingly met with a type of expertise based on subjective experiences that patients generate through these online platforms. Previous studies have deemed this expertise *experiential knowledge* (Bellander and Landqvist 2020; Versteeg et al. 2018) while other authors have referred to it as *embodied knowledge* (Ellingson 2006). In this context, experiential knowledge not only describes how patients experience the symptoms of a condition through their bodies (e.g., unknown adverse effects of drugs) but also incorporates information related to their experiences with the healthcare system (e.g., difficulties accessing a healthcare professional). Since this paper focuses on the non-knowledge related to bodily experiences, we will hereafter refer to the expertise that patients produce through social media as *embodied knowledge* (see also Parviainen and Aromaa 2017).

Before the emergence of these online platforms, embodied knowledge was usually discussed with close relatives, family, and friends. However, social media has allowed the distribution of this expertise also among networks of weak ties. Evidence indicates that online networks of weak ties better satisfy the informational and emotional support needs of patients, as the perspectives encountered are more diverse than those found in networks of strong ties (Wright 2016). In turn, these practices make patients feel empowered to share the expertise created online with their healthcare professionals. The problem remains that healthcare professionals mainly approach the treatment of a condition from the perspective of disease, meaning that these individuals have difficulties to manage embodied knowledge (Rosendal et al. 2017). That is, although healthcare professionals recognise the limits of their expert knowledge (Wilkesmann 2016), most are not prepared to deal with unknowns that cannot be objectively measured. Consequently, the use of social

media for health-related purposes is still controversial. A case study in Finland conducted with healthcare professionals is consistent with this argument (Torkkola et al. 2019). Of the 490 participants, 2 out of 10 admitted to experiencing conflict with patients over their social media practices. Most of these disputes were generated because the patient made expertise claims based on misleading information (Torkkola et al. 2019).

From an epistemological perspective, one way of framing the distinctive understandings of social media between patients and healthcare professionals could be through *the digital technology paradox*. As Mather and Cummings (2019, p. 1–2) explained, "there is an inability of health professionals […] to access digital technology in the workplace, while it is increasingly recognized that its use has the potential to improve patient outcomes". Regulatory restrictions are indeed one of the reasons for this misalignment (Harris et al. 2013). Nevertheless, the sociotechnical imaginaries of healthcare organisations could also be influencing the adoption of social media in the organisation and delivery of healthcare (Jasanoff and Kim 2009). Sociotechnical imaginaries "are generally future-oriented visions of connected social and technological orders, with more or less determinism built into them" (Sismondo 2020, p. 505). This means that the incorporation of social media into professional settings may depend, for example, on the evidence that healthcare organisations have regarding these online platforms (Flear 2019).

That is, if the sociotechnical imaginary of a healthcare organisation is built only from its experiences with social media, these online platforms will probably not be a part of its imagined future because the evidence available to it describes more harm than benefit. Previous studies identified some of these harms, which go from misinformation and disinformation to the possible commodification of the patient opinion (Lupton 2014; Torkkola et al. 2019). However, the reasons behind the lack of engagement on the part of healthcare professionals when it comes to social media seem to go beyond structural inertia. In terms of power, these online platforms have always generated concerns about the authority of healthcare professionals (Sendra and Farré 2017; Torkkola et al. 2019). Social media is helping patients to focus on aspects of a condition that were previously not acknowledged, thus generating a type of expertise based on non-knowledge (i.e., unknowns). Therefore, when patients present it during medical consultations, the power relationship changes, and healthcare professionals can no longer sustain their position of superiority over these individuals (Wagner et al. 2021). From this perspective, we argue that patients' embodied knowledge created through these online platforms can be described as healthcare professionals' non-knowledge.

Table 1. Healthcare professionals' management of embodied knowledge generated through social media.

Strategy	Definition	Application example
Denial	Uncomfortable knowledge is totally refused	Healthcare professionals do not acknowledge patients' knowledge generated through social media
Dismissal	It is admitted that uncomfortable knowledge exists, but then it is also refused for reasons such as unreliability or lack of precision	Healthcare professionals acknowledge patients' knowledge generated through social media, but the information is later discarded by citing misinformation concerns
Diversion	A distraction is created to get away from the uncomfortable knowledge	Healthcare professionals invite patients to consult other informational resources, such as scientific papers
Displacement	Distraction is created to avoid the uncomfortable knowledge and alternatives are placed	Informational resources recommended by healthcare professionals become alternatives to knowledge generated through social media

Source. Definitions paraphrased from Rayner (2012); examples provided by the authors.

If we understand non-knowledge as a space with power struggles (Perron et al. 2020), healthcare professionals could be framing social media as a technology that compromises their authority and highlights the limits of their expertise. Since embodied knowledge generated through these online platforms questions what healthcare professionals already know, this type of expertise becomes *uncomfortable knowledge* for healthcare professionals. Rayner (2012, p. 111) defines uncomfortable knowledge as any "potential information that presents either sort of danger to institutions". Uncomfortable knowledge can be managed through denial, dismissal, diversion, or displacement (Rayner 2012). Of these four categories, 'dismissal' seems to best describe the management of embodied knowledge generated through social media exercised by healthcare professionals (Table 1). For example, evidence indicates that healthcare professionals are aware of the practices that happen through these online platforms and its benefits for patients (Torkkola et al. 2019). However, instead of confronting their non-knowledge, embodied knowledge is often considered inferior (Flear 2019) and classified as medically unexplained symptoms (Rosendal et al. 2017). Consequently, in addition to perpetuate the prejudices that biomedicine holds against them (Sendra and Farré 2020), patients with conditions that cannot be proved with evidence (e.g.,

fibromyalgia) may end being victims of testimonial injustice (Blease et al. 2017; Fricker 2007).

3   NON-KNOWLEDGE AS A WAY TO UNDERSTAND THE DISMISSAL OF EMBODIED KNOWLEDGE

The literature on ignorance studies indicates that epistemic actors adopt different behaviours when it comes to manage non-knowledge, which can be positive (where non-knowledge is followed up) or negative (where non-knowledge is not considered) (Gross 2019). Some of these behaviours include *rational motivated ignorance* (Williams 2021a), *ignorance due to strong assumptions* (Stein 2020) and *strategic ignorance* (McGoey 2012, 2020). Although none of these behaviours has been studied in the context of the uses of social media in healthcare, they can still provide relevant insights on why healthcare professionals usually dismiss the embodied knowledge that patients generate through these online platforms. First, rational motivated ignorance describes situations where "individuals remain ignorant not because of a lack of available information, and not because of the various costs associated with acquiring that information, but because of the costs associated with knowledge itself" (Williams 2021a, p. 7823–7824). In this case, epistemic actors act from a perspective of self-protection, reaching the conclusion that it is better to avoid the unknown than to face their non-knowledge (Roberts 2013; Williams 2021b). According to this behaviour, healthcare professionals would be dismissing the embodied knowledge that arises from patients' social media practices to protect their position. This is consistent with the authority concerns previously described, as Williams (2021b) explained that confronting non-knowledge compromises the epistemic status of an individual.

Second, ignorance due to strong assumptions is related "to assuming an unequal distribution of human freedom" in decision-making processes, where the hierarchies within an organisation seem to have an influencing role (Stein 2020, p. 431). According to this behaviour, healthcare professionals would be dismissing the embodied knowledge that arises from patients' social media practices by assuming that knowledge production in medicine is their responsibility and not that of others. This lines up not only with the impact of these online platforms on medical power but also with the depth of organisational culture roots of healthcare organisations (Manning-Cork 2019). In this context, it could be argued that patients may also engage in this behaviour when they generate embodied knowledge. However, it is outside the scope of this paper to analyse this aspect in further detail. Third, strategic ignorance "illuminates the way various states and processes of unknowability are often structured by the power of some social groups to remain deliberately ignorant" (McGoey 2020, p. 198; McGoey 2012). According to this behaviour, healthcare professionals would be dismissing the embodied knowledge that arises from patients' social media practices to protect their organisations. This is because healthcare organisations should be understood as negotiated orders where

multiple, complex processes and services are intertwined (Tjora and Scambler 2009). Therefore, if patients that generate embodied knowledge through these online platforms were considered subjects with multiple epistemic positions instead of passive agents (Perron et al. 2020), the established order of healthcare organisations risks being destabilised.

4   CONSEQUENCES   OF   DISMISSING   EMBODIED KNOWLEDGE FOR HEALTHCARE

The distinctive understandings of social media between patients and healthcare professionals have both positive and negative connotations. On the one hand, these online platforms are beneficial because patients can share their embodied knowledge with other individuals, who also share their own, thereby forming a community of practice where lay expertise is exchanged between members (Wagner et al. 2019). These interactions create *epistemic value* (Barret et al. 2016), where patients perceive their conditions as authentic by highlighting issues that they were otherwise unable to see. People living with long COVID are the latest example of patients' social media practices. Indeed, Callard and Perego (2021) argued that long-term COVID-19 is the first disease to be defined using these online platforms. Despite a lack of expert knowledge, patients turned to social media to find evidence for embodied knowledge around their experiences, look for answers to their symptoms and claim testimonial. While these processes are similar for other conditions (Sendra and Farré 2020), the difference in the case of people living with long COVID lies in how quickly healthcare professionals confronted their non-knowledge (Callard and Perego 2021).

This precedent shows how patients make use of social media to make the unknown known. First, by sharing their experiences through these online platforms, social media assisted people living with long COVID to transform the 'unknown unknowns' of this condition into 'known unknowns'. Second, once the unknowns about long-term COVID-19 became a 'known unknown', the capacity of these online platforms for connection allowed patients to reach one another, slowly creating an epistemic community (Akrich 2010) that attracted the attention of experts. For healthcare professionals, the tipping point between knowledge and non-knowledge came when they began to seriously consider the experiences of people living with long COVID posted in social media. In other words, only when healthcare professionals confronted their non-knowledge, the 'known unknowns' of this condition converted into 'known knowns' through scientific research. Therefore, these online platforms offer an opportunity to improve the work practices of healthcare professionals, as patients with conditions that are invisible to evidence-based medicine will continue to use social media to manage their conditions. Embodied knowledge generated through these online platforms could also represent a change for rare diseases, where the "scarcity of expertise poses a

huge challenge to patients who seek access to diagnostic testing and appropriate treatment" (Dawkins et al. 2018, p. 12).

On the other hand, the risk with social media when it comes to embodied knowledge is that not all embodied knowledge is constructed equally. Previous research in the context of chronic pain suggested that patients undergo a three-step process when posting their experiences in social media, deciding in one of these steps which is the best way to share their expertise (Sendra and Farré 2020). Maslen and Lupton (2019, p. 1638) added that patient enactments on these online platforms may include "1) expertise claims based on appropriation and distribution of scientific knowledge and experience; 2) sharing experiential knowledge without claiming expertise and 3) evaluation and use of knowledge presented by others, principally through observing". Other studies have identified similar knowledge construction processes in blogs and forums (Bellander and Landqvist 2020; Versteeg et al. 2018). The problem remains that embodied knowledge may be generated from misleading information (Torkkola et al. 2019). In this context, previous research suggested that expertise claims based on misleading/incomplete information may end in patient disaffection with their healthcare professionals (Bellander and Landqvist 2020). We argue that dismissing embodied knowledge that arises from patients' social media practices not only generates tensions between patients and healthcare professionals, but it also creates problems with digitisation processes and the confidence that the public has in expert knowledge (Table 2).

Table 2. Harms related to patients' and healthcare professionals' distinctive understandings of social media.

Level	Actors involved	Potential harm
Macro	Society in general	Decreased public trust in expert knowledge
Meso	Healthcare organisations	Problems integrating innovations and consolidating patient-centred care
Micro	Patients and healthcare professionals	Increased tensions in relationships between patients and healthcare professionals

At the micro level, the distinctive understandings of social media between patients and healthcare professionals may increase tensions in their relationship. As described above, misleading information can damage trust between patients and healthcare professionals (Bellander and Landqvist 2020), which in turn causes the latter to become more guarded when it comes to their relationship (Laurent-Simpson and Lo 2019; Nie et al. 2018). The problem remains that trust is one of the key resources that healthcare professionals have at their disposal to tackle misleading information, especially when constructed using bottom-up approaches (Rodgers and Massac 2020). Previous research also indicates that the phenomenon of misleading information seems to be linked with a general disaffection towards information provided by traditional sources such as news media (Nielsen and

Graves 2017). Therefore, healthcare professionals need to confront embodied knowledge that arises from patients' social media practices not only to improve their communication with patients but also increase public trust in expert knowledge (Van Dijck and Alinead 2020). Otherwise, avoiding or overlooking this risk may lead to disaster (Erikainen et al. 2019; Perron et al. 2020), such as the 'infodemic' surrounding the COVID-19 pandemic (Van Dijck and Alinead 2020).

At the meso level, the distinctive understandings of social media between patients and healthcare professionals complicates both digitisation processes and the consolidation of patient-centred care. While these online platforms have made healthcare more participatory (Sosnowy 2014), biomedical evidence often remains the only information that healthcare professionals consider when making decisions. These deep-rooted practices may be related to micro- and macro-ignorance (McGoey 2020), with the latter corresponding to the organisational posture that results from incorporating individual positions into a situation. For example, if most healthcare professionals still perceive social media as negative (micro-ignorance), these beliefs will add up and reinforce the position that these online platforms are harmful for their work practices (macro-ignorance). As Laurent-Simpson and Lo (2019, p. 1283) explained, "understanding and respecting what drives the personal perspectives reflected on social media pages should better position medical professionals and other scientists to win buy-in from the audience". Therefore, healthcare professionals need to confront embodied knowledge that arises from patients' social media practices to learn how this expertise can benefit their work practices (Nerland and Hasu 2021). Otherwise, failing to consider this harm will continue to hinder the transition towards a (fully) patient-centred, technology-based healthcare.

At the macro level, the distinctive understandings of social media between patients and healthcare professionals may decrease public trust in expert knowledge. For example, misleading information may lead to epistemological populism, concept that describes "the favouring of 'common people's knowledge' over knowledge that is produced by expert systems" (Numerato et al. 2019, p. 84). As indicated elsewhere, social media has the potential to exacerbate epistemological populism (Numerato et al. 2019), partly due to the echo chamber phenomenon linked to these platforms (Usher and Ng 2020). While epistemological populism goes beyond embodied knowledge, it could be argued that social media 'echo chambers' where only patients take part are a potential starting point for this phenomenon. Similarly, healthcare professionals should be aware that online communities are not only diverse and work in different ways (Usher and Ng 2020), but also that patients may have different epistemic statuses within these groups (Tempini and Del Savio 2019). Therefore, healthcare professionals need to confront embodied knowledge that arises from patients' social media practices to understand how to adapt their strategies to the multiple realities of these individuals (Larson 2018; Timmermann 2020; Will 2020). Otherwise, disregarding this harm could increase the campaigns based on misinformation that already take place on

these online platforms, such as the anti-vaccination movement (Hernandez et al. 2021).

5   TOWARDS THE MANAGEMENT OF THE UNKNOWN

As we argued throughout the paper, there is a need for healthcare professionals to confront the embodied knowledge that arises from patients' social media practices. The problem remains that these online platforms and other digital health technologies have created professionalism-related issues for healthcare professionals. Some of these problems involve uncertainty surrounding the concepts of authenticity, subjectification, and diversity when health professionals use avatars (Hallqvist 2019) or perceiving as inappropriate the use of smartphones while interacting with patients (Naples et al. 2020). As for social media, evidence indicates healthcare professionals' difficulties discerning what constitutes professional versus unprofessional behaviour on these online platforms (Ahmed et al. 2020; Curtis and Gillen 2019; Ruan et al. 2020) or how to balance the use of social media between personal and professional lives (Jones et al. 2021; Ventola 2014). Therefore, confronting the embodied knowledge that arises from patients' social media practices goes beyond achieving an organisational effort to frame these online platforms as epistemic objects that can complement expert knowledge and improve the organisation and delivery of healthcare.

To be able to manage patients' social media practices, healthcare professionals also need to improve their digital health literacy (Bimber and Gil de Zúñiga 2020). As indicated elsewhere, healthcare professionals are not immune to problems related to misleading information (Palomino-Gonzales et al. 2020). A way of addressing this issue is by improving the education of healthcare professionals on digitization (Mather and Cummings 2019). To this end, some medical undergraduate programs have started to introduce digital professionalism into their syllabuses, though this is still uncommon (Ahmed et al. 2020). The aim of this type of professionalism is to help healthcare professionals learn how to make sense of digital technologies in relation to their work practices (Mather and Cummings 2019). Previous research has found that digital professionalism helps medical students consider both the positive and negative aspects of social media (Jones et al. 2021). Therefore, introducing these courses would not only help healthcare professionals address the mistrust present in their relationship with patients but they would also have more resources to deal with misleading information (Nie et al. 2018; Ruan et al. 2020).

In turn, digital professionalism would allow healthcare professionals to enhance the digital health literacy of their patients (Palomino-Gonzales et al. 2020). This is significant considering that the challenge of the COVID-19 pandemic has been "not to keep people informed so much as to help them navigate the sheer volume of novel information and claims made available everyday" (Marchal and Au 2020, p. 3). Other studies have also indicated that it is equally

important to improve the level of media literacy of the public (Pulido et al. 2020; Ratzan et al. 2020), especially when expert knowledge is increasingly biomediatized (Briggs and Hallin 2016). That is, people need not only the skills to process the information found online but also tools that will allow them to determine the reliability of the source (Pulido et al. 2020; Ratzan et al. 2020). For these courses to be successful, evidence also suggests that healthcare organisations must establish them at the same time as digital health technologies are introduced into the work practices of healthcare professionals (Mather and Cummings 2019). The distinctive understandings of social media between patients and healthcare professionals discussed throughout this paper demonstrates how this simultaneous application has not yet happened in the case of these online platforms.

Similarly, healthcare professionals should be more open-minded in relation to digitisation and embodied knowledge, especially when their beliefs seem to be a crucial element to transform the 'known unknowns' of social media into 'known knowns'. In this context, healthcare professionals must be not only motivated to confront their non-knowledge but also interested in the spaces that generate it (Nerland and Hasu 2021; Wilkesmann 2016). This would give healthcare professionals the opportunity to expand the limits of their expert knowledge (Roberts 2013) while also developing closer relationships with patients, which in turn would facilitate the delivery of patient-centred care. As Piras and Miele (2019, p. 128) explained, communication between patients and healthcare professionals mediated through technology "allows personal events, previously not taken into account, to be communicated and considered by providers". However, digitisation also comes with the risk of patients being more responsible for their conditions than the healthcare professionals who care for them (Erikainen et al. 2019). While healthcare organisations use 'responsibilisation' as a way of controlling risks, this shift in responsibility could increase inequalities and create new ways of medicalisation (Erikainen et al. 2019; Hofmann and Svenaeus 2018; Jasper 2020).

Lastly, the distinctive understandings of social media between patients and healthcare professionals are related not only to the attitude of healthcare professionals but also to the allocation of communication responsibilities within healthcare organisations. The use of these online platforms in hospitals is generally handled by media specialists in communication departments, where the strategies are mostly constructed from an institutional perspective (Costa-Sánchez and Míguez-González 2018). Allocating communication tasks this way may generate tensions between managerial points of view and healthcare professionals' perspectives. Hospitals are not only negotiated orders (Tjora and Scambler 2009) but also embodied informational structures, where the integration of a technology cannot be treated exclusively as a material phenomenon (Parviainen and Koski 2023). For example, consider messages from patients shared on the Facebook page of a hospital where they discuss their symptoms. Healthcare professionals will not address these messages because they believe that it is not their responsibility to answer them. Their current communication duties do not go beyond interpersonal

communication at the micro level. In turn, as communication specialists work under this corporate framework, they also may not respond to these messages because they do not know how to address them, perpetuating the distinctive understandings of social media between patients and healthcare professionals.

6   CONCLUSION

Starting from an epistemological framework, this paper has problematised the distinctive understandings of social media between patients and healthcare professionals from a different perspective. The paper has illustrated the ways that non-knowledge influences healthcare professionals when managing embodied knowledge that patients generate through social media. It has also been argued that patients are increasingly forming epistemic communities on these online platforms while healthcare professionals seem to be engaging in various strategies to protect their position of power and avoid facing the unknowns that arise from patients' social media practices. Likewise, it has been discussed that the distinctive understandings of these online platforms between patients and healthcare professionals are potentially harming the organisation and delivery of healthcare at all levels (Table 2), including fostering trust issues in their relationship. The paper has culminated by suggesting that initiatives aimed at increasing media and digital health literacy skills of both patients and healthcare professionals could be one approach for resolving these issues. Regardless of this problematization, what seems clear is that the current level of circulation of misleading information on social media necessitates some sort of action from healthcare professionals (Pulido et al. 2020; Van Dijck and Alinead 2020).

However, some limitations should also be considered. As stated in the introduction, the paper used only secondary sources. Carrying out a more detailed study would require the collection of completely new material, where medical ignorance would be the focus of the research questions. For this reason, we can only speculate that non-knowledge is a potential barrier to the successful digitisation of healthcare. Future studies should explore this topic through other methodologies, such as semi-structured interviews with healthcare professionals or focus groups with policy regulators. Similarly, the paper has discussed the processes of knowledge-production from a broad perspective without considering that it takes time for knowledge to become evidence-based or that patients may lack part of the specialized vocabulary related to their conditions. While discussing these issues are beyond the reach of this paper, future research should consider the processes of knowledge-production in further detail. Another limitation is related to the field of ignorance studies itself, which is a relatively new discipline within epistemology (Gross and McGoey 2015; 2022). This means that existing research may not entirely capture the complexity of the topic discussed in this paper. Despite these shortcomings, the ideas introduced here open a range of possibilities to approach the study of digitisation from a new perspective. We have shown that patients' social

media practices can become another epistemic focus for healthcare professionals to supplement their expert knowledge. Embodied knowledge is becoming stronger in these online platforms, and the COVID-19 pandemic is a clear example of how healthcare professionals should address, sooner rather than later, lay expertise generated through social media (Van Dijck and Alinead 2020). Otherwise, in addition to perpetuating epistemic injustices (Blease et al. 2017; Flear 2019), healthcare professionals will continue to experience problems digitising healthcare.

FUNDING STATEMENT AND ACKNOWLEDGMENTS

COMPETING INTERESTS

No potential conflict of interest was reported by the author(s). An abstract of this paper was presented at the NordMedia 2021 (virtual) conference.

ORCID

Anna Sendra: http://orcid.org/0000-0001-7069-1378
Sinikka Torkkola: https://orcid.org/0000-0002-1150-6627
Jaana Parviainen: https://orcid.org/0000-0002-7838-592X

AUTHOR CONTRIBUTIONS

Anna Sendra: conceptualisation; writing – original draft; writing – review & editing. Sinikka Torkkola: conceptualisation; supervision; writing – review & editing. Jaana Parviainen: conceptualisation; supervision; writing – review & editing.

REFERENCES

Ahmed, W., Jagsi, R., Gutheil, T.G., and M.S. Katz. 2020. Public disclosure on social media of identifiable patient information by health professionals: Content analysis of Twitter data. *Journal of Medical Internet Research* 22 (9): e19746. https://doi.org/10.2196/19746

Akrich, M. 2010. From communities of practice to epistemic communities: Health mobilizations on the Internet. *Sociological Research Online* 15 (2): 1–17. https://doi.org/10.5153/sro.2152

Barrett, M., Oborn, E., and W. Orlikowski. 2016. Creating value in online communities: The sociomaterial configuring of strategy, platform, and stakeholder engagement.

Information Systems Research 27 (4): 704–723. https://doi.org/10.1287/isre.2016.0648

Bellander, T., and M. Lanqvist. 2020. Becoming the expert constructing health knowledge in epistemic communities online. *Information, Communication & Society* 23 (4): 507–522. https://doi.org/10.1080/1369118X.2018.1518474

Benetoli, A., Chen, T.F., and P. Aslani. 2018. How patients' use of social media impacts their interactions with healthcare professionals. *Patient Education and Counseling* 101 (3): 439–444. https://doi.org/10.1016/j.pec.2017.08.015

Bimber, B., and H. Gil de Zúñiga. 2020. The unedited public sphere. *New Media & Society* 22 (4): 700–715. https://doi.org/10.1177/1461444819893980

Blease, C., Carel, H., and K. Geraghty. 2017. Epistemic injustice in healthcare encounters: Evidence from chronic fatigue syndrome. *Journal of Medical Ethics* 43 (8): 549–557. http://dx.doi.org/10.1136/medethics-2016-103691

Briggs, C.L., and D.C. Hallin. 2016. *Making health public: How news coverage is remaking media, medicine, and contemporary life*. London and New York: Routledge.

Callard, F., and E. Perego. 2021. How and why patients made Long Covid. *Social Science & Medicine* 268: 113426. https://doi.org/10.1016/j.socscimed.2020.113426

Costa-Sánchez, C., and M. Míguez-González. 2018. Use of social media for health education and corporate communication of hospitals. *El Profesional de la Información* 27 (5): 1145–1154. https://doi.org/10.3145/epi.2018.sep.18

Curtis, F., and J. Gillen. 2019. "I don't see myself as a 40-year-old on Facebook": Medical students' dilemmas in developing professionalism with social media. *Journal of Further and Higher Education* 43 (2): 251–262. https://doi.org/10.1080/0309877X.2017.1359503

Dawkins, H.J.S., Draghia-Akli, R., Lasko, P., Lau, L.P.L., Jonker, A.H., Cutillo, C.M., Rath, A., Boycott, K.M., Baynam, G., Lochmüller, H., Kaufmann, P., Le Cam, Y., Hivert, V., Austin, C.P., and International Rare Diseases Research Consortium (IRDiRC). 2018. Progress in rare diseases research 2010–2016: An IRDiRC perspective. *Clinical and Translational Science* 11 (1): 11–20. https://doi.org/10.1111/cts.12501

El Kassar, N. 2018. What ignorance really is: Examining the foundations of epistemology of ignorance. *Social Epistemology* 32 (5): 300–310. https://doi.org/10.1080/02691728.2018.1518498

Ellingson, L.L. 2006. Embodied knowledge: Writing researchers' bodies into qualitative health research. *Qualitative Health Research* 16 (2): 298–310. https://doi.org/10.1177/1049732305281944

Erikainen, S., Pickersgill, M., Cunningham-Burley, S., and S. Chan. 2019. Patienthood and participation in the digital era. *Digital Health* 5: 2055207619845546. https://doi.org/10.1177/2055207619845546

Flear, M.L. 2019. Epistemic injustice as a basis for failure? Health research regulation, technological risk and the foundations of harm and its prevention. *European Journal of Risk Regulation* 10 (4): 693–721. https://doi.org/10.1017/err.2019.67

Fricker, M. 2007. *Epistemic injustice: Power & the ethics of knowing*. Oxford: Oxford University Press.

Fuller, S. 2002. *Social epistemology* (2nd ed.). Bloomington: Indiana University Press.

Gross, M. 2019. Not Knowing as Luxury: Strategic Nonknowledge and the Demand for a "Sportbrake". *Luxury* 6 (1): 63–81. https://doi.org/10.1080/20511817.2018.1738705

Gross, M., and L. McGoey. 2015. Introduction. In *Routledge international handbook of ignorance studies*, ed. M. Gross and L. McGoey, 1–14. London and New York: Routledge.

Gross, M., and L. McGoey. 2022. Revolutionary epistemology: the promise and peril of ignorance studies. In *Routledge international handbook of ignorance studies* (2nd ed.), ed. M. Gross and L. McGoey, 3–14. London and New York: Routledge.

Hallqvist, J. 2019. Digital health and the embodying of professionalism: Avatars as health professionals in Sweden. *Professions and Professionalism* 9 (2): e2847. https://doi.org/10.7577/pp.2847

Harris, J.K., Mueller, N.L., and D. Snider. 2013. Social media adoption in local health departments nationwide. *American Journal of Public Health* 103 (9): 1700–1707. https://doi.org/10.2105/AJPH.2012.301166

Hernandez, R.G., Hagen, L., Walker, K., O'Leary, H., and C. Lengacher. 2021. The COVID-19 vaccine social media infodemic: Healthcare providers' missed dose in addressing misinformation and vaccine hesitancy. *Human Vaccines & Immunotherapeutics* 17 (9): 2962–2964. https://doi.org/10.1080/21645515.2021.1912551

Hofmann, B., and F. Svenaeus. 2018. How medical technologies shape the experience of illness. *Life Sciences, Society and Policy* 14 (1): 3. https://doi.org/10.1186/s40504-018-0069-y

Jasanoff, S., and S. Kim. 2009. Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva* 47 (2): 119–146. https://doi.org/10.1007/s11024-009-9124-4

Jasper, U. 2020. The anticipative medicalization of life: Governing future risk and uncertainty in (global) health. In *The politics and science of prevision: Governing and probing the future*, ed. A. Wenger, U. Jasper and M.D. Cavelty, 122–140. London: Routledge.

Jones, S., Chudleigh, M., Baines, R., and R.B. Jones. 2021. Did introducing Twitter and digital professionalism as an assessed element of the nursing curriculum impact social media related incidence of 'Fitness to practise': 12-year case review. *Nurse Education* in Practice 50: 102950. https://doi.org/10.1016/j.nepr.2020.102950

Kerwin, A. 1993. None too solid: Medical ignorance. *Knowledge* 15 (2): 166–185. https://doi.org/10.1177/107554709301500204

Killbride, M.K., and S. Joffe. 2018. The new age of patient autonomy: Implications for the patient-physician relationship. *JAMA: The Journal of the American Medical Association* 320 (19): 1973–1974. https://doi.org/10.1001/jama.2018.14382

Larson, H.J. 2018. The biggest pandemic risk? Viral misinformation. *Nature* 562 (7727): 309. https://doi.org/10.1038/d41586-018-07034-4

Laurent-Simpson, A., and C.C. Lo. 2019. Risk society online: Zika virus, social media and distrust in the Centers for Disease Control and Prevention. *Sociology of Health & Illness* 41 (7): 1270–1288. https://doi.org/10.1111/1467-9566.12924

Lupton, D. 2014. The commodification of patient opinion: The digital patient experience economy in the age of big data. *Sociology of Health & Illness* 36 (6): 856–869. https://doi.org/10.1111/1467-9566.12109

Lupton, D. 2018. *Digital health: Critical and cross-disciplinary perspectives*. London: Routledge.

Manning-Cork, N. 2019. Advancing professionalism, attentive to culture, to improve health systems. *Medical Education* 53 (11): 1069–1071. https://doi.org/10.1111/medu.13960

Marchal, N., and H. Au. 2020. "Coronavirus EXPLAINED": YouTube, COVID-19, and the socio-technical mediation of expertise. *Social Media + Society* 6 (3): 205630512094815. https://doi.org/10.1177/2056305120948158

Martin, G.P., Currie, G., and R. Finn. 2009. Reconfiguring or reproducing intra-professional boundaries? Specialist expertise, generalist knowledge and the 'modernization' of the medical workforce. *Social Science & Medicine* 68 (7): 1191–1198. https://doi.org/10.1016/j.socscimed.2009.01.006

Maslen, S., and D. Lupton. 2019. 'Keeping it real': Women's enactments of lay health knowledges and expertise on Facebook. *Sociology of Health & Illness* 41 (8): 1637–1651. https://doi.org/10.1111/1467-9566.12982

Mather, C.A., and E. Cummings. 2019. Developing and sustaining digital professionalism: A model for assessing readiness of healthcare environments and capability of nurses. *BMJ Health & Care Informatics* 26 (1): e100062. http://dx.doi.org/10.1136/bmjhci-2019-100062

McGoey, L. 2012. The logic of strategic ignorance. *The British Journal of Sociology* 63 (3): 533–576. https://doi.org/10.1111/j.1468-4446.2012.01424.x

McGoey, L. 2020. Micro-ignorance and macro-ignorance in the social sciences. *Social Research* 87 (1): 197–217. https://doi.org/10.1353/sor.2020.0014

McPherson, S., Rost, F., Sidhu, S., and M. Dennis. 2020. Non-strategic ignorance: Considering the potential for a paradigm shift in evidence-based mental health. *Health* 24 (1): 3–20. https://doi.org/10.1177/1363459318785720

Naples, R., Costas-Chavarri, A., Golden, D.W., Gmitter, E., French, J.C., and J.M. Lipman. 2020. Digital professionalism in patient care: A case-based survey of surgery faculty and trainees. *The Journal of Surgical Research* 253: 193–200. https://doi.org/10.1016/j.jss.2020.03.057

Nerland, M., and M. Hasu. 2021. Challenging the belief in simple solutions: The need for epistemic practices in professional work. *Medical Education* 55 (1): 65–71. https://doi.org/10.1111/medu.14294

Nie, J., Cheng, Y., Zou, X., Gong, N., Tucker, J.D., Wong, B., and A. Kleinman. 2018. The vicious circle of patient–physician mistrust in China: Health professionals' perspectives, institutional conflict of interest, and building trust through medical professionalism. *Developing World Bioethics* 18 (1): 26–36. https://doi.org/10.1111/dewb.12170

Nielsen, R.K., and L. Graves. 2017. "News you don't believe": Audience perspectives on fake news. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-10/Nielsen%26Graves_factsheet_1710v3_FINAL_download.pdf. Accessed 21 June 2021.

Numerato, D., Vochocová, L., Štětka, V., and A. Macková. 2019. The vaccination debate in the "post-truth" era: Social media as sites of multi-layered reflexivity. *Sociology of Health & Illness* 41 (S1): 82–97. https://doi.org/10.1111/1467-9566.12873

Palomino-Gonzales, M.M., Lovón-Cueva, M.A., and R.d.C. Arellanos-Tafur. 2020. The health network and its participation in the dissemination or containment of fake news and hoaxes related to COVID-19: The case of Lima-Peru. Chasqui: *Revista Latinoamericana De Comunicación* 1 (145): 93–118. https://revistachasqui.org/index.php/chasqui/article/view/4332

Parviainen, J., and J. Aromaa. 2017. Bodily knowledge beyond motor skills and physical fitness: A phenomenological description of knowledge formation in physical training. *Sport, Education and Society* 22 (4): 477–492. https://doi.org/10.1080/13573322.2015.1054273

Parviainen, J., and A. Koski. 2023. 'In the future, as robots become more widespread': A phenomenological approach to imaginary technologies in healthcare organisations. In *The Oxford Handbook of Phenomenologies and Organization Studies*, ed. F.X. De Vaujany, J. Aroles and M. Perézts, 277–296. Oxford: Oxford University Press.

Perron, A., Rudge, T., and M. Gagnon. 2020. Hypervisible nurses: Effects of circulating ignorance and knowledge on acts of whistleblowing in health. *Advances in Nursing Science* 43 (2): 114–131. https://doi.org/10.1097/ANS.0000000000000311

Piras, E.M., and F. Miele. 2019. On digital intimacy: Redefining provider–patient relationships in remote monitoring. *Sociology of Health & Illness* 41: 116–131. https://doi.org/10.1111/1467-9566.12947

Pulido, C.M., Ruiz-Eugenio, L., Redondo-Sama, G., and B. Villarejo-Carballido. 2020. A new application of social impact in social media for overcoming fake news in health. *International Journal of Environmental Research and Public Health* 17 (7): 2430. https://doi.org/10.3390/ijerph17072430

Ratzan, S.C., Sommariva, S., and L. Rauh. 2020. Enhancing global health communication during a crisis: Lessons from the COVID-19 pandemic. *Public Health Research & Practice* 30 (2): 3022010. https://doi.org/10.17061/phrp3022010

Rayner, S. 2012. Uncomfortable knowledge: The social construction of ignorance in science and environmental policy discourses. *Economy and Society* 41 (1): 107–125. https://doi.org/10.1080/03085147.2011.637335

Roberts, J. 2013. Organizational ignorance: Towards a managerial perspective on the unknown. *Management Learning* 44 (3): 215–236. https://doi.org/10.1177/1350507612443208

Rodgers, K., and N. Massac. 2020. Misinformation: A threat to the public's health and the public health system. *Journal of Public Health Management and Practice* 26 (3): 294–296. https://doi.org/10.1097/PHH.0000000000001163

Rosendal, M., Olde Hartman, T.C., Aamland, A., Van Der Horst, H., Lucassen, P., Budtz-Lilly, A., and C. Burton. 2017. "Medically unexplained" symptoms and symptom disorders in primary care: Prognosis-based recognition and classification. *BMC Family Practice* 18: 18. https://doi.org/10.1186/s12875-017-0592-6

Ruan, B., Yilmaz, Y., Lu, D., Lee, M., and T.M. Chan. 2020. Defining the digital self: A qualitative study to explore the digital component of professional identity in the health professions. *Journal of Medical Internet Research* 22 (9): e21416. https://doi.org/10.2196/21416

Sendra, A., and J. Farré. 2017. Institutional pain communication via Twitter by Spanish and US pain societies: Analysis of levels of use and engagement. *Catalan Journal of Communication & Cultural Studies* 9 (1): 3–23. https://doi.org/10.1386/cjcs.9.1.3_1

Sendra, A., and J. Farré. 2020. Communicating the experience of chronic pain through social media: Patients' narrative practices on Instagram. *Journal of Communication in Healthcare* 13 (1): 46–54. https://doi.org/10.1080/17538068.2020.1752982

Sendra, A., Farré, J., and R.W. Vaagan. 2020. Seeking, sharing and co-creating: A systematic review of the relation between social support theory, social media use and chronic diseases. *Social Theory & Health* 18 (4): 317–339. https://doi.org/10.1057/s41285-019-00106-z

Sismondo, S. 2020. Sociotechnical imaginaries: An accidental themed issue. *Social Studies of Science* 50 (4): 505–507. https://doi.org/10.1177/0306312720944753

Sosnowy, C. 2014. Practicing patienthood online: Social media, chronic illness, and lay expertise. *Societies* 4 (2): 316–329. https://doi.org/10.3390/soc4020316

Stein, F. 2020. Blinded by the slide show: Ignorance and the commodification of expertise. *Critique of Anthropology* 40 (4): 420–437. https://doi.org/10.1177/0308275X20959409

Tempini, N., and L. Del Savio. 2019. Digital orphans: Data closure and openness in patient-powered networks. *BioSocieties* 14 (2): 205–227. https://doi.org/10.1057/s41292-018-0125-0

Timmermann, C. 2020. Epistemic ignorance, poverty and the COVID-19 pandemic. *Asian Bioethics Review* 12 (4): 519–527. https://doi.org/10.1007/s41649-020-00140-4

Tjora, A.H., and G. Scambler. 2009. Square pegs in round holes: Information systems, hospitals and the significance of contextual awareness. *Social Science & Medicine* 68 (3): 519–525. https://doi.org/10.1016/j.socscimed.2008.11.005

Torkkola, S., Vuolanto, P., and J. Parviainen. 2019. *Social media and professional communication in the health care organization*. NordMedia 2019: Communication, Creativity & Imagination – Challenging the Field, 21–23 August, Malmö.

Usher, N., and Y.M.M. Ng. 2020. Sharing knowledge and "Microbubbles": Epistemic communities and insularity in US political journalism. *Social Media + Society* 6 (2): 205630512092663. https://doi.org/10.1177/2056305120926639

Van Dijck, J., and D. Alinead. 2020. Social media and trust in scientific expertise: Debating the COVID-19 pandemic in the Netherlands. *Social Media + Society* 6 (4). https://doi.org/10.1177/2056305120981057

Ventola, C.L. 2014. Social media and health care professionals: Benefits, risks, and best practices. *Pharmacy and Therapeutics* 39 (7): 491–520. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103576/

Versteeg, W., Te Molder, H., and P. Sneijder. 2018. "Listen to your body": Participants' alternative to science in online health discussions. *Health* 22 (5): 432–450. https://doi.org/10.1177/1363459317695632

Wagner, A., Polak, P., and M. Świątkiewicz-Mośny. 2019. Who defines – Who decides? Theorising the epistemic communities, communities of practice and interest groups in the healthcare field: A discursive approach. *Social Theory & Health* 17 (2): 192–212. https://doi.org/10.1057/s41285-018-0073-6

Wagner, A., Polak, P., and M. Świątkiewicz-Mośny. 2021. From community of practice to epistemic community – Law, discipline and security in the battle for the legalisation of medical cannabis in Poland. *Sociology of Health & Illness* 43 (2): 316–335. https://doi.org/10.1111/1467-9566.13217

Wehling, P. 2015. Fighting a losing battle? The right not to know and the dynamics of biomedical knowledge production. In *Routledge international handbook of ignorance studies*, ed. M. Gross and L. McGoey, 206–214. London and New York: Routledge.

Whooley, O., and K.K. Barker. 2021. Uncertain and under quarantine: Toward a sociology of medical ignorance. *Journal of Health and Social Behavior* 62 (3): 271–285. https://doi.org/10.1177/00221465211009202

Wilkesmann, M. 2016. Ignorance management in hospitals. *VINE Journal of Information and Knowledge Management Systems* 46 (4): 430–449. https://doi.org/10.1108/VJIKMS-08-2016-0046

Will, C.M. 2020. The problem and the productivity of ignorance: Public health campaigns on antibiotic stewardship. *The Sociological Review* 68 (1): 55–76. https://doi.org/10.1177/0038026119887330

Williams, D. 2021a. Motivated ignorance, rationality, and democratic politics. *Synthese* 198: 7807–7827. https://doi.org/10.1007/s11229-020-02549-8

Williams, D. 2021b. To communicate scientific research, we need to confront motivated ignorance. LSE Impact Blog. https://blogs.lse.ac.uk/impactofsocialsciences/2021/01/13/to-communicate-scientific-research-we-need-to-confront-motivated-ignorance/. Accessed 21 June 2021.

Wright, K.B. 2016. Communication in health-related online social support groups/communities: A review of research on predictors of participation, applications of social support theory, and health outcomes. *Review of Communication Research* 4: 65–87. https://doi.org/10.12840/issn.2255-4165.2016.04.01.010

CENTRALISING QUALITATIVE RESEARCH IN BIG DATA METHODS THROUGH ALGORITHMIC ETHNOGRAPHY

Naomi Barnes[a] and Sam Hames[b]

ABSTRACT

Responding to the challenge for qualitative researchers to claim a central place in conversations about big data, analytics, datafication, data mining and the role of algorithms, this article describes a mixed-method research partnership focused on algorithmic ethnography. In the debates about the opacity of online algorithms, qualitative researchers typically advocate for access to code. This standard discourse centralises the technical aspects of big data and networked ethnographies. Instead, this article outlines a research methodology that analyses algorithmic discourses by working alongside the technical expertise of data scientists and utilizes the affordability of big data methods to do qualitative work. The potential for qualitative research skills to investigate the underlying technical processes that frame online social interactions is proposed as a way to place how people understand the world at the centre of big data research.

Keywords: big data; algorithms; ethnography; algorithmic discourses; mixed methods.

[a] Queensland University of Technology, Australia
[b] University of Queensland, Australia.

1   INTRODUCTION

Big data and their subsequent networks have historically been associated with qualitative research (Bancroft et al., 2014); however, the introduction of Internet-mediated big data has meant the methodologies have been taken over by quantitative, statistics-oriented, technology-enhanced methodologies (Sarkar, 2021). As such, qualitative research is faced with the challenge of re-centralising itself in this revolution of how "developing technology *and* the way people interface with that technology in their social contexts *and* what that interface enables or leads to in those contexts" (Cheek, 2021, p. 124). To rise to this challenge, the following article describes a research approach developed through a mixed-method partnership where the goal was to centralise qualitative research using contemporary big data technologies. In particular, we[1] focused on developing a technique to "measure the implicit meanings that occur in-between strings of words" (Mills, 2018, p. 599) by analysing algorithmic discourses.

This article has two purposes. Firstly, to explain a qualitative digital methodological approach that emerged through discussions between an education sociologist and a data scientist. By working together to understand the language of each other's fields, we outline below the first steps we have taken in re-centring big data towards qualitative, interpretive analysis. Secondly, we piloted what we have developed with a field of study— political and policy discourses associated with education. In particular, we look at the digital rhetoric that formed around a literacy policy deliberation process in Australia. Unlike an empirical research report where the methodology proceeds the findings, in this paper we weave the story of our field of study into the methodological explanation for illustrative purposes. You can read about the empirical study and findings in more detail elsewhere (Barnes, 2021). Before proceeding we explain the interpretive foundation to our study of the effects of algorithms, digital rhetoric.

2   DIGITAL RHETORIC

It has become increasingly important to consider the role of algorithms in the discourses influencing politics. A decade of research into the "black boxed" effects of algorithms (Pasquale, 2015) have led online communications scholars to argue that the opacity of algorithms are a key sociotechnical problem that requires transparency and regulation. Social algorithm researchers who have studied their effects have insisted that algorithms be opened up (Eubanks, 2018; O'Neil, 2016; Pasquale, 2020) for true social critique. As such, the direction of research has moved further away from the sociological and humanities approaches that have traditionally dominated an understanding of social issues and interactions, towards

---

[1] While each of us brought very different but complementary skills to the project, for ease of explanation, we use the pronoun "we" throughout, rather than specifically indicating how tasks were split.

technical approaches that required specialised computing skills (Mills, 2018). However, as algorithms are vastly differentiated, constantly evolving through machine learning, and intersecting across multiple platforms, with "long chains of actors, technologies and meanings" (Christin, 2020b, p. 897), it seemed reasonable to us to explore how qualitative approaches, well versed in collecting and analysing dynamic data, could be combined with specialist technical expertise to better understand the effect of algorithms.

To develop this methodology, we drew on the field of digital rhetoric. According to Eyman (2015) the term *digital rhetoric* is perhaps most simply defined as 'the application of rhetorical theory (as analytic method or heuristic for production) to digital texts and performances' (p.45). Some approaches are closely related to traditional rhetoric and composition studies including, how people use strategies to analyse digital texts, identifying how digital texts are constructed in order to produce more effective communication objects, and how people create digital authorship identities and audiences. These are all important to our project but are well established in qualitative studies and the leap from analysing terrestrial texts, audience and authorship to digital versions is not what we are concerned with in this article. Instead we are interested in explaining a methodology for considering the 'rhetorical function of networks' (Eyman, 2015, p. 45) by concentrating on a key mechanism for holding those networks together — algorithms.

When a political entity wishes to influence, then knowledge of how algorithms deliver that information becomes a rhetorical tool of influence. Education policy and politics is the field we chose to pilot our algorithmic rhetorical analysis. As education is a political field, it is then important to consider the role of algorithmic discourses in how education policy is developed and enacted. Close attention has been placed on algorithms in educational research in the fields that would be expected, such as the increasing reliance on machine learning and automated decision making in using data to construct educational futures (see for example Webb et al., 2020) and the calculation of A-Level results in the UK and Ireland (Kelly, 2021). As educational policy is developed in the public sphere, we contend that the discursive effect of algorithms through public-facing websites and applications, such as social media, are also important to consider. We hypothesised that educational rhetoric could be identified through manipulation of big data networking capabilities and forensic examination of the education and policy actors' rhetoric but also the algorithmic mechanisms that connected those policy actors. Unlocking algorithms' effect on educational political rhetoric is a broader project our collaboration is working towards and how we illustrate the methodological approach we explain in this article.

2.1 Digital rhetoric and algorithmic ethnography

Algorithmic ethnography is the approach we took to understanding how algorithms hold networks together. Our approach involved analysing what information was fed

into the so-called black box of the Internet and theorizing, using algorithmic metaphors, the educational discourses which emerge. Algorithmic ethnography has recently been defined as the "ethnographic study of the computational systems enabling and shaping online interactions" (Christin, 2020a, p. 109). Situated within the realm of online communication, Christin's research approach has the potential to be expanded into a sociological systems approach that considers the rhetorical role of algorithms in shaping online and offline interactions. By combining two methodological elements of Christin's (2020a, 2020b) proposal to enrol algorithms in established digital ethnographic approaches, and digital rhetoric (Losh, 2009), this paper outlines a methodology for analysing effects of algorithms in online educational discourse.

According to Christin (2020a), "adopting the lens of algorithmic ethnography entails paying close attention to the role of algorithms in structuring the back and front end of the digital platforms that increasingly mediate digital exchanges" (p. 109). While researchers like Pasquale (2015) advocate for making transparent the mechanisms which deliver information, algorithmic ethnography provides a way to theorise what is happening behind the forward-facing text to boost or block how that text is distributed around the Internet. Without the fine detail in the code which platforms keep black-boxed, it is still possible to hypothesise about the digital rhetoric of the algorithms. There are a finite number of categories of algorithm which means there are a finite number of potential interpretations for how information is being distributed online. We use Christian and Griffiths' (2017) popular explanation[2] (drawing non-exhaustively from many subfields) as a starting framework for our algorithmic thinking, with a particular focus on sorting and caching. We use these algorithms as building blocks to help us bridge the space between qualitative inquiry and the quantitative worldview underpinning the more complex algorithmic assemblages deployed in online systems.

Christin (2020a) outlines three necessary steps for designing a qualitative algorithmic ethnography: the type of data collected, the role of algorithms in sorting and organizing the data, and the effects of online metrics on how people interact online. To develop hypotheses about the role of algorithms we revisited a recent education policy deliberation study (Barnes, 2021). The research question we were interested in was: *What role can we see algorithms playing in affecting how online users influence online literacy policy deliberation?* Considering the so-called Reading Wars (see for example Pearson, 2004), we determined it was a useful place

---

[2] We have selected this popular guide to algorithms, rather than an academic text as 1) the book gives a clear explanation of how each algorithmic category works with accessible scenarios. We have not critically engaged with the work because 2) this methodological approach is just beginning, and we would hope other researchers will find a way into discovering the effects of the algorithms we did not note. Furthermore, 3) there is not enough room in this paper to effectively describe all the categories, so a popular explanatory text is a good place to direct anyone interested in pursuing this methodological approach. We would hope that more critical sociological work could emerge from this starting point.

to begin looking closely at the digital rhetoric. In brief, the Reading Wars are the academic, political, and public debate about the best way to teach children reading. Today the Science of Reading (Castles et al., 2018) has been determined by educational policymakers to be the best evidence-based approach. When we conducted this research, the Science of Reading had not yet gained policy status and the debate we analysed was part of the political process by which advocates of the program advocated for the program. The online engagement we captured through algorithmic ethnographic methods were between what we will refer to as Science of Reading (SOR) advocates versus socio-cultural literacy advocates (SCL). Very basically, SCL practitioners advocate for reading to be taught in the context of books, while SOR practitioners advocate for reading to initially be taught out of context through repetition using objects like flashcards and drills. This choice was also made because the Reading Wars have existed before the Internet was invented, meaning that the historical manifestations of the debate could be used to make sense of the debate as it occurred online.

Our method uses three analytical phases and one theoretical phase – we will start with a high-level overview of these phases to frame the detailed case study that follows. Of course, although we present these phases as a linear ordering for the convenience of the reader this is only an approximation of the actual reality of conducting such an analysis.

2.1.1   Phase 1: Selection of data

First it is necessary to determine which data is to be included in the study – this is also necessarily the first point of qualitative interpretation of the data. While this phase may initially begin with simple computational filtering, such as selection of documents containing a keyword, the selection phase would iteratively move towards more and more qualitative decision making about relevance of individual data points. Additionally, it is at this stage that the unit of analysis (that is - what is a data point?) is also chosen.

2.1.2   Phase 2: Sorting and searching through the data

Having defined what data is to be included in the analysis, the next phase delved further into the qualitative interpretation of the data. At this point we used the logic of an algorithmic cache to conceptualise the data coding procedure in a way that is consistent with the computational requirements of the next phase.

2.1.3   Phase 3: Conceptualising the social network

The third analytical phase uses network visualisations to map out the contours of the searched and sorted data. This phase brought together the different concepts identified by searching and sorting through the constructed caches of phase 2, along with the data selected as part of phase 1 into a single unified view.

2.1.4 Phase 4: Hypothesising the algorithmic discourses

Connecting the conceptualisations of the network to how the algorithm affected online rhetoric was determined through abductive reasoning. This phase did not intend to contribute to existing scholarship through deductive or inductive outcomes, rather develop hypotheses through experimenting with visualisations and various ways of analysing the data. The hypotheses form the basis of research questions for future inductive and deductive analysis framed by a relevant sociological theory.

2.2 Phase 1: Selection of data

The selection of data involved a collaboration between the authors: our first author being the educational sociologist intent on understanding online educational policy advocacy; our second author being a research data scientist. The collaboration began with a feasibility conversation, trying to find the middle ground of what our first author wanted to research and what our second author was able to extract from the available databases. This process comprised of discussions about research questions, how the extraction process worked, and what was available for extraction. At the time of data extraction, the Australian Twitter database was the most comprehensive social media database available for use. Today other databases are also possible; however, the database had multiple holes in time that data were not archived, so a phenomenon needed to be selected that aligned with the available timeframes. After some initial searches, the literacy policy debate, that became known as the 2018 #PhonicsDebate, was identified as a viable study for two reasons. First, it was selected because audience members were encouraged to tweet using the hashtag #PhonicsDebate during a live debate on YouTube meaning there would be a lot of tweets within a short timeframe. Secondly, it became a multiplatform (Twitter, Facebook, YouTube, blogs, podcasts, petitions, static websites, in person debates) event requiring qualitative skills to connect them all together. Quantitative data collection can only connect between platforms via similarity of key words. Meaning needs interpretive skills that software cannot yet do. This choice meant that qualitative work was centralised.

The key to centralising qualitative research in this research approach was that although this initial starting point is still a quantitative content-based selection, it was conducted in the context of a quantitative/qualitative collaboration with clear parameters and mutual discussion of goals. This created a solid starting point for further qualitative refinement and exploration outside of those parameters.

It was also at this stage that we decided on the initial ethical framework for approaching this data. In doing so we considered multiple aspects, including the public (and publicised) nature of the debate, the technically public nature of the tweets, the nature of the expected audience authoring those tweets, and the difficulty of paraphrasing or other types of anonymisation when reporting on social

media data. Weighing up these considerations, we decided that when reporting on this material would not identify or use material on individual participants, not even in paraphrased or "anonymised" forms – all reporting would focus on aggregated or abstracted views of data, and high-level descriptions of content that cannot be linked to individual accounts. This study was also approved by the QUT Human Research Ethics Committee.

2.2.1 Extraction

A spreadsheet (mocked up[3] in Table 1) was provided, drawn from QUT's Digital Observatory's longitudinal Australian Twittersphere database using #phonicsdebate, #phonicscheck, associated key words "teaching +reading", and "phonics". These were collected from the two-week period surrounding either side of 31st July 2018 when a debate about the value of universal synthetic phonics was live streamed on YouTube. The back and forth between us continued once the phenomenon and timeframe was selected, refining the inclusion and exclusion criteria. Hames engaged in computational activities like pruning keywords that were overly broad or were capturing unrelated tweets.

Table 1

Mockup of the initial extraction of tweet data, including initial fields considered, and the structural data describing the tweets place in the Twitter conversation.

tweet_id	user	text	created_at	reply	retweet	quote
XXXXXX177 1462070000	@usera	Phonics is the best way to teach reading! #phonicsdebate	26/07/2018 7:23	0	0	0
XXXXXX485 1394900000	@userb	Phonics should not be separate to language #phonicsdebate	26/07/2018 8:15	0	0	0
XXXXXX785 6229600000	@userc	@userf Don't teachers usually do both? #phonicsdebate	26/07/2018 9:46	1	0	0
XXXXXX249 9598490000	@userd	RT @usere After the #phonicsdebate, come over to #AussieEd!	26/07/2018 11:24	0	1	0

---

[3] As the ethical clearance does not allow for the direct quotation of tweets, the tables are illustrations of the EXCEL sheets. The tweet identification numbers have been anonymized to ensure the original tweets are not searchable.

The unit of analysis was the tweet (or the text written by Twitter users to comment on the phenomenon), and the associated Twitter users *handle* (online unique identifier), and associated metadata (tweet id and time tweeted). The tweet was chosen as the unit of analysis because this aligns the qualitative and computational components of this work directly with the fundamental unit of communication on Twitter.

Alternative units of analysis were considered, including user and conversation focused approaches. A user focused approach would consider the unit of analysis to be a user profile and an aggregation of all of their tweets, representing a particular accounts communications relating to the subject of interest. A conversation focused unit of analysis would aggregate users and tweets replying to a specific thread into a single unit, representing a specific exchange relating to the topic of interest - at the time of this case study the Twitter API did not provide the information needed to ensure that complete threads of conversation could be reconstructed, ruling out this approach. Importantly aspects of the user and conversation focused units can be aggregated from the tweet level representation, but the disaggregation to the tweet level is more difficult.

Once the tweets were computationally identified, the associated information attached to each tweet was also extracted, such as who wrote the tweet and what time it was broadcast. The tweets were those from identified Australian Twitter accounts, but the dataset did include international participation if an Australian account retweeted a tweet by an international user. Tweet extraction did not include tweets or profile information from Twitter accounts that were 'protected', that is, those whose tweets and profile details are accessible only to other users approved by the account holder.

2.2.2   Refinement and initial exploration

The refinement and exploration process are the key qualitative approaches within the selection phase. The refinement process involved a manual reading of each of the tweets to ensure that they were all part of the phenomenon under scrutiny. What became evident was that the key words "teaching +reading" extracted multiple tweets that were unrelated to the policy debate or education all together. These tweets were removed from the spreadsheet. The final list of tweets was refined from N=2232 to N=2150 tweets.

The initial exploration also revealed that a number of tweets were repeated in the list because each time a user retweeted a tweet by another user that was already in the dataset, that retweet also became a part of that users' timeline of tweets (effectively, if a tweet was retweeted ten times in the dataset, it would appear as ten distinct rows in the spreadsheet, despite having the same content). A qualitative decision was made to leave them as separate tweets rather than collapse them into one tweet with a retweet count. This work could have been done computationally,

but we decided that because the research was a sociology it was important to retain the interactions between users to better understand the social dynamics[4].

The initial exploration also allowed for a qualitative coding of the tweeters' experiences of the phenomenon. By keeping the primary research question in mind (What are the social dynamics of the online literacy policy pipeline?), tweets were roughly coded using qualitative decision making about how each tweeter experienced, understood, comprehended and/or conceptualized the phonics debate. For example, some tweeters were positively on one side or the other, some were diplomatic, and some were using the opportunity to use the hashtag for promotion of other online educational events and hashtag chats (See table 2 as a mockup of this exploration).

Table 2

Explorative coding of tweets. The tweet_id, user and Text columns are as shown in Table 1, and indicate the "as collected" state of the data, the remaining four columns indicate a binary coding (code is present or absent) of the associated tweet. The developed categories identify the position of the tweet with respect to the sides of the Debate.

tweet_id	user	Text	SCL	SOR	Diplomatic	Marketing
XXXXXX 1771462 070000	@usera	Phonics is the best way to teach reading! #phonicsdebate	0	1	0	0
XXXXXX 4851394 900000	@userb	Phonics should not be separate to language #phonicsdebate	1	0	0	0
XXXXXX 7856229 600000	@userc	@userf Don't teachers usually do both? #phonicsdebate	0	0	1	0
XXXXXX 2499598 490000	@userd	RT @usere After the #phonicsdebate, come over to #AussieEd!	0	0	0	1

This coding prepared the data for further qualitative targeted coding, later in the analysis by making chunks of data sortable. This binary coding scheme is commonly

---

[4] If a similar study were to be conducted again, the list of "likes" would also have been useful to look at the social dynamics of phonics debate sentiment (which users have liked which tweets only became comprehensibly accessible via the Twitter API as of early 2022 and was not possible at the time of the study).

used for statistical and machine learning representations of data for its efficient numerical representation – for the qualitative use case here it uses the affordances of spreadsheets for data entry, while allowing easier integration of the coded data into other analytical tools.

2.3 Phase 2: Searching and sorting

Searching and sorting are often used interchangeably in regard to the Internet, as "search engines" are actually sorting engines. For example, Google uses personal data to *sort* the millions of websites and deliver suggestions in response to a *search* term. In the second analytical phase, these processes are two separate but overlapping activities aimed at developing rigorous categories and themes that could be used to describe the experiences of the phonics debate.

2.3.1 Key terms, hashtags, and time

While the tweets had been extracted via umbrella terms, like #phonicsdebate, other keywords and phrases were evident in the tweets. These keywords gave clues about the themes and categories that could be extracted from the data. Tableau was used to quantitatively sort through the tweets and identify key terms which indicated which experiences of the phonics debate required deeper qualitative analysis. For example, in analysing the secondary hashtags (see Figure 1), "dyslexia", "science" or "evidence" might be clues for sorting the tweets into themes.
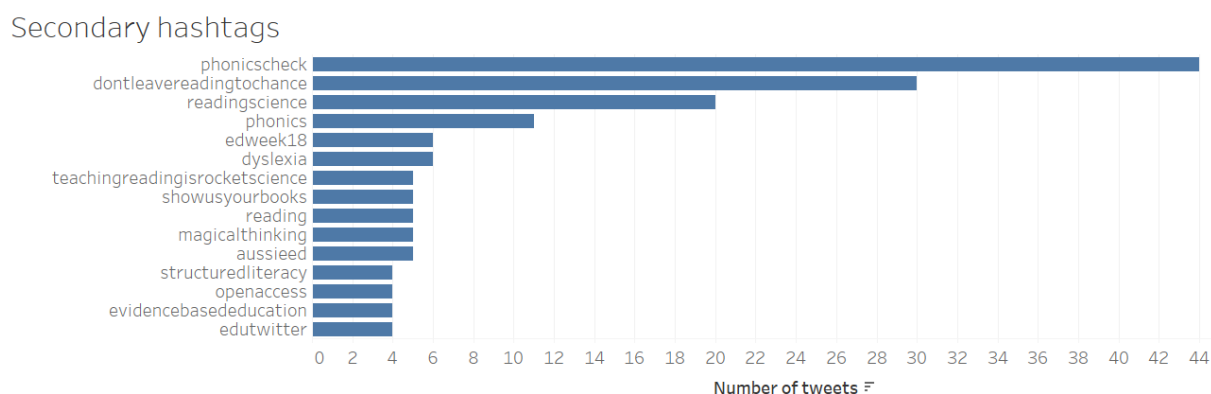

Figure 1. An example of sorting secondary hashtags in the twitter data by frequency to complement the close reading with existing context about the data.

Other key term analysis software, like Leximancer or Excel, could also be used to deal with the quantity of tweets, but Tableau was chosen because it could be used to build a visualisation of the data contemporaneously, rather than after, the coding of the data. As the point of the research was to experiment with what different

visualisations large amounts of data could produce, such a tool was more useful for developing hypotheses than the others on offer.

Tableau also allowed for counting the number of tweets broadcast by users over the course of the debate, which allowed judgement about intensity of sentiment, or potential bot engagement[5] with the hashtag. For example, the most visible user (who we later referred to as the hyper-connector below) retweeted multiple tweets (N=476) from the SOR side of the debate. Basic analysis of the tweets from the most visible users showed that parents of children with dyslexia were the most engaged groups of actors in the dataset.

2.3.2   Qualitative caching

The reason we have termed this stage "caching" is because the technical structure of the Internet uses caching to speed up an individual's access to information. In computing, a cache is a copy of some data stored in a temporary (usually ephemeral) location for either easy access or to avoid repeating an intensive operation. A physical analogue would be the sorting trolley in a library before books are returned to the stacks. The caches of information also work to define what a term will come to mean for each Internet user. For example, when you search for a term on Google and explore the initial offerings, Google's search engine will have a cache of websites you clicked through to in case you would like to visit the website again. In other words, Google's search engine begins to build a personalised "meaning" of a search term for each user with each website they visit that uses those terms.

We felt this action is a good description of the initial sorting of themes when coding qualitative research. For example, when a qualitative researcher sorts their sticky notes, decides on codes for NVivo, or, in our case, uses a binary coding system to organize themes, the most recent piece of data added to a pile is the pathway into the clearest definition of the theme. When a qualitative researcher cannot choose a pile or *cache*, they will either manipulate the definition of an already existing cache or start a new cache. Either way, the most recent piece of data is the clearest clue for the defining features of the cache.

Unlike computational caching where the first pieces of information eventually drop out of the cache in a purely mechanical process, *qualitative caching* is an iterative process which occurs throughout all analytical phases. A qualitative researcher will return to all the pieces of data in a cache in order to develop a definition of that theme, discarding or shifting data points between caches, until all relevant pieces of data have a home, and the qualitative researcher is satisfied that all the pieces of data within a cache are representative of the theme. In this framework the qualitative choices an analyst makes are: 1) which items are worth including in the cache for further consideration and 2) in which cache (or caches) should they be included.

---

[5] No bots were detected in this study after investigating the top tweeters.

The binary coding (1 and 0) in our project was used to develop the caches and iteratively revisit them until satisfied with the themes. A binary code means the Excel spread sheet could be sorted and resorted and checked and rechecked each time a cache meaning shifted. This also meant that the close reading and rigorous coding could be done in chunks, rather than having to read the entire spreadsheet every time a new concept was noted. Considering that a debate is a structured genre where points are made, illustrated, and rebutted, it was straight forward to align tweets with the argumentation. Not all tweets engaged directly with the live debate, and those were put aside for future sorting (see Table 3).

Table 3

Argumentation coding of tweets – this extends on Table 2 to show the additional binary codes used for the argumentation analysis. The additional columns in Table 2 are omitted for brevity.

tweet_id	user	Text	Claim	Warrant	Rebuttal
XXXXXX 1771462070000	@usera	Phonics is the best way to teach reading! #phonicsdebate	1	0	0
XXXXXX 4851394900000	@userb	Phonics should not be separate to language #phonicsdebate	0	1	0
XXXXXX 7856229600000	@userc	@userf Don't teachers usually do both? #phonicsdebate	0	0	1

Each cache organised the vast number of initial tweets into manageable sized groupings for later forensic qualitative analysis.

2.4 Phase 3: Conceptualizing the network

The searching and sorting phase of data analysis provided the foundation for interpreting the social network analysis and conceptualising the digital rhetoric. The #phonicsdebate social network analysis was used to render an initial network visualisation, which helped us analyse that network and subsequently used the analysis to begin to explore variations within the network.

2.4.1 Mapping the network

The Twitter data was also rendered for a social network analysis using the open-source tool Gephi and its Force Atlas 2 algorithm to lay out the network. Interpreting what was computed by this algorithm (see Figure 2) showed that there was a distinct binary between each side of the debate.
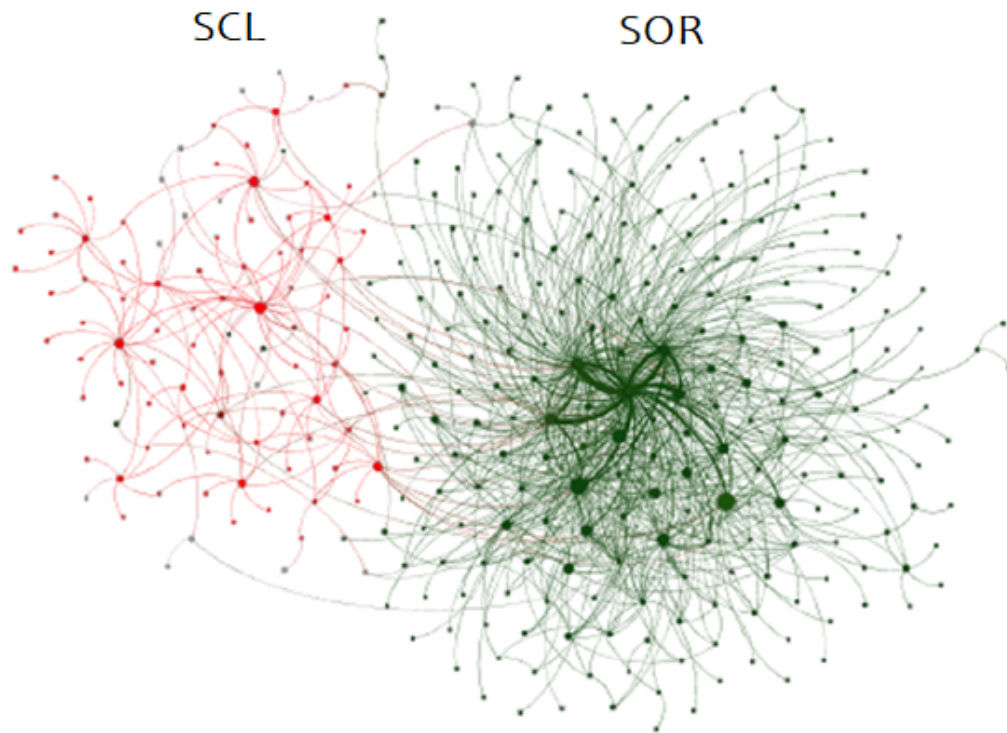
Figure 2. Initial social network analysis. Each node in this network is a Twitter account, the edges between nodes indicate the volume of engagement (retweets and replies) between accounts – thicker edges indicate stronger engagement. Nodes are coloured using the sentiment assigned to each profile to indicate positioning of that account with respect to the debate (red - SCL, green – SOR, grey - diplomats and marketers).

Each node was given a colour according to the side of the debate their tweets indicated they supported. One challenge for this process was that the qualitative analysis was conducted at the granularity of the tweet, but this visualisation was created with the nodes as users. In other words, to colour the nodes, the tweets needed to be read and interpreted because the position of the individual users was in what they wrote in their tweets (refer back to Table 3). Such coding of data would not be possible with quantitative analysis only. This manual annotation was also only possible due to the relatively small number of nodes. This process would not have been infeasible for a larger or more complex dataset and how to address this issue for larger datasets is the focus of our ongoing collaboration.

We noted that the SOR side of the debate had a hyper-connector, or someone who was using the functions of Twitter to distribute a huge amount to tweets from the SOR perspective. This was the same parent who retweeted 476 times. In Figure 2, this hyper-connector is the node in the centre of the SOR network surrounded by a daisy shape. This daisy means they were actively tweeting, retweeting and replying to multiple accounts. Moreso that anyone else but not alone

in their activity as is indicated by the thicker connecting lines (or edges) within the network visualisation.

2.4.2 Variation of experience

Much analytical work will stop at the visualization of a dataset, but our process began after all the possible visualisations were developed from the raw data. In other words, the visuals helped us make sense of a massive amount of data to begin hypothesising about the role of algorithms in distributing information on the Internet. The initial social network analysis produced a stripped version of the phenomenon (See Figure 2). It showed that there were two sides of the literacy debate and that they were quite obviously on two ends of a spectrum. However, the social network analysis raised questions which sparked further investigation: These included:

- Who is the hyper-connector and is their activity driving the movement of information around Twitter or are they being assisted in any way by other groups?

- Why are there so few people tweeting about one perspective and so many about the other?

- Is it an accurate representation of online engagement with the debate?

The third question led to a re-coding of the data to create a simple tweak to the network map. In the first and second phases of the analysis, we noticed the tweet texts did not always directly link into the debaters. For various reasons that range from privacy to non-existent Twitter accounts, some debate participants were named but not coded in a way that automatic coding could render them into a social network. Automatic coding needed each member of the network to have a consistent username. As such, we searched the tweets for variants and adjusted the text for consistency (see Table 4) so that the automated systems would identify all the interlocutors.

Table 4

Inconsistent username examples

Hypothetical Twitter account	Possible Variants of Name and title (could include misspelling)
@ProfessorX	Professor X, Prof X, PX, Dr X, Speaker 1, X, Professor Ex, etc
@HarleyQuinn	Professor Quinn, Prof Quinn, Dr Quinn, Harly Quin, Dr Q, Q, HQ, etc

This process transformed the initial network representing only the concrete, platform mediated traces of communication to a more nuanced network that incorporates references to people, not just social media accounts. In the resulting social network one side of the debate all but disappeared from the network map, being absorbed into the other (see Figure 3).
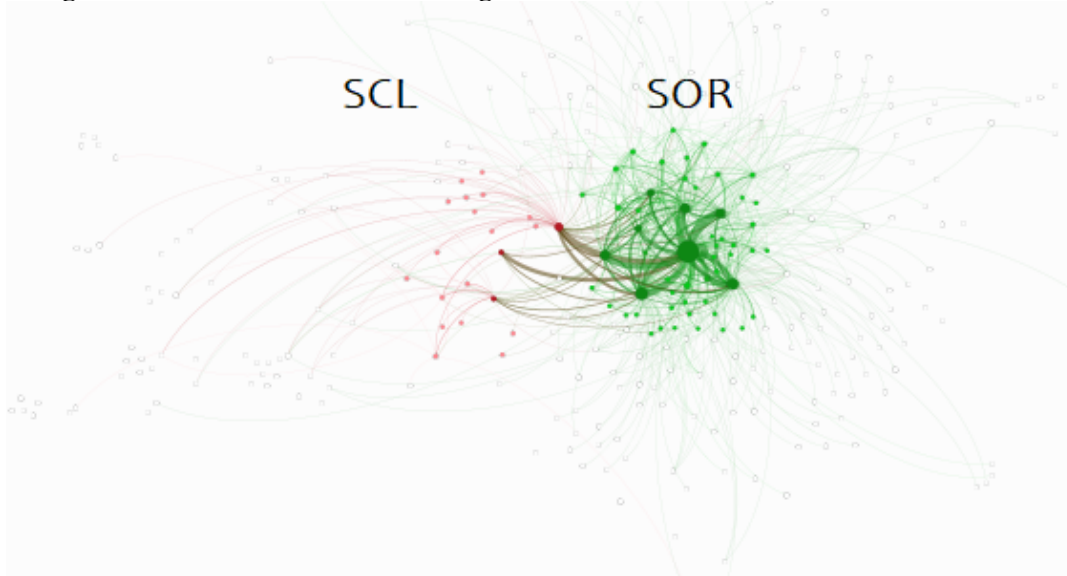

Figure 3. Manipulated social network – As in Figure 2, each node in this network is a Twitter account, the edges between nodes indicate the volume of engagement (retweets and replies) however this time edge weights include mentions of debate participants by name rather than just by Twitter handle. Thicker edges indicate stronger engagement. Nodes are coloured using the sentiment assigned to each profile to indicate positioning of that account with respect to the debate (red - SCL, green -SOR, uncoloured – diplomats and marketers).

This raised new questions about how representative the debate was on Twitter. The reconfiguration of the social network analysis showed that the same accounts were still hyper-connectors, initiating an investigation into why their Twitter behaviour worked so powerfully on the network.

Hypothesising the different experiences of the phonics debate from different interlocutors within the debate provided a clue for where to search for data next. Conceivably, the extra data might have come from any source, including interviews, but as the purpose of the research was to discover how information about a policy moves through the Internet, we chose to stay with data available in Internet archives. However, this article is about centralising qualitative research in big data research and how we worked together to conceptualise how data science and digital sociology can work together. As such we have chosen to not explain this later forensic stage, but it can be read about elsewhere (Barnes, 2021). Essentially, analysis of other online objects led us to return to the Twitter data set and more

closely consider the connection between the tweets and the debate. What became apparent was that the hyper-connectors and the bloggers were seeing the phonics debate from a <u>parent's perspective</u>. Knowing that parents were the key hyper-connectors we began to theorise what was rhetorically happening behind the forward-facing text.

Hypothesis building: When a child finds it difficult to read, the response from the parent (or family system) is very different from the response of the school system or the literacy research system. Coming to understand how a parent might use the platform is one research question we drew from the data (as outlined in the next section).

2.5 Phase 4: Hypothesising the algorithmic discourses

In this final phase of the process, we walk through the process of hypothesising using the outputs of the analytical phase and our algorithmic metaphors to arrive at a narrative description and map of the logics at work. We provide this as more of an extended example because this phase is expected to be the most specific and sensitive to the particular research questions of each project.

The initial hypothesising was informed by the algorithmic sorting digital rhetoric. Sorting algorithms, including the commonly used *Mergesort,* alongside caching algorithms, consider the optimal organization of information on the Internet. Christian and Griffiths (2017) provide the recursive logarithmic pattern behind the process and explain why it was so revolutionary by comparing it to how a human might organize their personal library. In terms of Mergesort, a practical and near-optimal way to organize a bookshelf is to invite friends around and divide the books evenly between them. Each friend is asked to organize their own stack, then stacks are combined – because each stack is already sorted combining them is easier than trying to sort everything at once. Using this concept, we can consider algorithmic organisation on the Internet as a collaborative effort – but rather than individual decisions about specific orders, we see individuals curate an organised view of their corner of platforms implicitly via the logic of what they consume and engage with. Platforms, attempting to "personalise" content then mediate the final merging of individual stacks not only based on personal use of the Internet but also the groups of users one might engage with the most — friends. While a human might sort their information into alphabetical order or by genre, theme, or topic, to access it at a later date, an algorithm deployed on the Internet will generally sort information using the logic behind caching. Algorithmically, caching is the most efficient way to find information. Sorting things into categories is less efficient that creating stacks of recently used information. The computer algorithm uses the logic that someone is more likely to want information closely related to the information they just consumed and engaged with. As such the quickest way to organize information is via a logic which directs the human to the last piece of information they used.

Adding the Mergesort logic to caching algorithms, an Internet site or application that sees its major function as searching (whether explicitly user directed or not) will recommend information closely related to the last piece of information extracted and connected to the groups a user is most likely to interact with.

In terms of the phonics debate study, we hypothesised that if a parent were to search Google, Facebook, or another social media site for why their child cannot read, they are very likely to come across other parents experiencing the same issues. Those parents then share their links with each other. These links are most likely to have been established before children attend school and be authored by health professionals like psychologists and speech therapists. Effectively the Mergesort friendship group grows but remains constrained to the presentation of information that aligns with what has already been seen. Eventually a parent may end up on Twitter, which has a strong teacher presence, and encounter the Reading Wars. However, by the time a parent arrives at the Reading Wars, they are more likely to side with the debate that is closer to their Mergesort friends – the psychologists and speech therapists. From a traditional rhetorical point of view, the parents have a logic of affect and comradery attached to their argument because of the Mergesort friends they gathered before coming across an alternative point of view.

3   REFLECTION

While our analysis has shown the more simplified algorithmic discourses (sorting and caching) are at work, we believe that there is enough evidence to justify future algorithmic ethnographies considering the role of algorithms in digital rhetoric. This is particularly important given the increasing role machine learning and artificial intelligence is taking in decision-making. Although this initial work has focused on two foundational classes of algorithms as organisational tools, algorithms as deployed in the real world can be much more complex – sorting and caching are much more likely to be used as building blocks of larger systems. Despite this limitation of this work, we think using these algorithmic tools to inform our analysis and theorising is a useful for ensuring that the qualitative and quantitative components of such work can be mutually grounding rather than separate.

Through our exploration of the possibilities of big and small data network analysis, we have shown that algorithmic ethnography that includes algorithmic digital rhetorical analysis, is a useful way forward in centralising qualitative research in big data methods. It should be noted that the dataset presented here is small enough to work forensically with each node, edge, and network representation. Larger datasets will require more comprehensive qualitative and quantitative efforts: for this study off the shelf tools and simple data formats worked, but "scaling up" to map out a larger phenomenon will require more detailed attention to the modelling of data, the representation of qualitative labelling efforts and how the components are drawn together into the final map. This study is a first step in how

mixed-methods teams can work together, with the purpose of understanding each researcher's field enough to solve such problems.

As moderation and connection of the different platforms became too unwieldly and enormous to be conducted via human labour, algorithms became the vehicles responsible for doing the work, becoming increasingly sophisticated by applying machine learning to distribute information and users more quickly and efficiently around the Internet. Algorithms became the lifeblood of the Internet, and increasingly tangled and rooted in how people navigate information about society, including education, and use that knowledge to make decisions. Now digital platforms linked to, but separate from, the political system, are woven into how and why political decision making is performed. In this reality transdisciplinarity becomes essential for understanding the effects of the Internet on policy deliberation and politics. There are too many systems at play for one qualitative researcher to come to understand and those systems are too dynamic for one data scientist to adequately capture.

4   REFERENCES

Bancroft, A., Karels, M., Murray, Ó. M., & Zimpfer, J. (2014). Not Being There: Research at a Distance with Video, Text and Speech. In *Big Data? Qualitative Approaches to Digital Research* (Vol. 13, pp. 137–153). Emerald Group Publishing Limited. https://doi.org/10.1108/S1042-319220140000013009

Barnes, N. (2021). The social life of literacy education: How the 2018 #phonicsdebate is reshaping the field. *The Australian Educational Researcher*. https://doi.org/10.1007/s13384-021-00451-x

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Cheek, J. (2021). Big Data, Thick Data, Digital Transformation, and the Fourth Industrial Revolution: Why Qualitative Inquiry Is More Relevant than Ever. In *Collaborative Futures in Qualitative Inquiry*. Routledge.

Christian, B., & Griffiths, T. (2017). *Algorithms to Live By: The Computer Science of Human Decisions*. HarperCollins GB.

Christin, A. (2020a). Algorithmic ethnography, during and after COVID-19. *Communication and the Public*, *5*(3–4), 108–111. https://doi.org/10.1177/2057047320959850

Christin, A. (2020b). The ethnographer and the algorithm: Beyond the black box. *Theory and Society*, *49*(5), 897–918. https://doi.org/10.1007/s11186-020-09411-3

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

Eyman, D. (2015). *Digital Rhetoric: Theory, Method, Practice*. University of Michigan Press.

Kelly, A. (2021). A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*, *47*(3), 725–741. https://doi.org/10.1002/berj.3705

Losh, E. M. (2009). *Virtualpolitik: An electronic history of government media-making in a time of war, scandal, disaster, miscommunication, and mistakes*. MIT Press.

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? *Qualitative Research*, *18*(6), 591–603. https://doi.org/10.1177/1468794117743465

O'Neil, C. (2016). *Weapons of Math Destruction*. Crown Publishing Group.

Pasquale, F. (2015). *The Black Box Society*. Harvard University Press.

Pasquale, F. (2020). *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Harvard University Press.

Pearson, P. D. (2004). The Reading Wars. *Educational Policy*, *18*(1), 216–252. https://doi.org/10.1177/0895904803260041

Sarkar, S. (2021). Using qualitative approaches in the era of big data: A confessional tale of a behavioral researcher. *Journal of Information Technology Case and Application Research*, *23*(2), 139–144. https://doi.org/10.1080/15228053.2021.1916229

Webb, P. T., Sellar, S., & Gulson, K. N. (2020). Anticipating education: Governing habits, memories and policy-futures. *Learning, Media and Technology*, *45*(3), 284–297. https://doi.org/10.1080/17439884.2020.1686015