

JOURNAL^{OF} DIGITAL SOCIAL RESEARCH

VOL 6 : NO 4 : 2024

Special issue

(UN)STABLE DIFFUSIONS THE PUBLICS, PUBLICITIES, AND PUBLICIZATIONS OF GENERATIVE AI

Guest edited by:

Fenwick McKelvey, Joanna Redden, Jonathan Roberge & Luke Stark

(UN)STABLE DIFFUSIONS: THE PUBLICS, PUBLICITIES, AND PUBLICIZATIONS OF GENERATIVE AI

Fenwick McKelvey, Joanna Redden, Jonathan Roberge & Luke Stark

"A.I. IS HOLDING A MIRROR TO OUR SOCIETY": LENSA AND THE DISCOURSE OF VISUAL GENERATIVE AI

Kate M. Miltner

AUTOMATED HIRING PLATFORMS AS MYTHMAKING MACHINES AND THEIR SYMBOLIC ECONOMY

Etienne Grenier & Nicholas Chartier-Edwards

A SIGN THAT SPELLS: MACHINIC CONCEPTS AND THE RACIAL POLITICS OF GENERATIVE AI

Fabian Offert & Thao Phan

THE TECHNOLOGICAL DRAMA OF AI: FROM PRIVATE POWER TO PLAYER ENGAGEMENT WITH ALPHAGO

Xerxes Minocher

AI POLICYMAKING AS DRAMA: STAGES, ROLES, AND GHOSTS IN AI GOVERNANCE IN THE UNITED KINGDOM AND CANADA

Alison Powell & Fenwick McKelvey

STRATEGIC MISRECOGNITION AND SPECULATIVE RITUALS IN GENERATIVE AI

Sun-ha Hong

GETTING DEMOCRACY WRONG: HOW LESSONS FROM BIOTECHNOLOGY CAN ILLUMINATE LIMITS OF THE ASIOMAR AI PRINCIPLES

Gwendolyn Blue & Mél Hogan

X-GRAM AND/AS TECHSPSURE: SPELLING OUT THE CLIMATE CONSEQUENCES OF GENERATIVE AI

Crystal Chokshi

JOURNAL^{OF} DIGITAL SOCIAL RESEARCH

VOL 6 : NO 4 : 2024

EDITORIAL BOARD

EDITOR-IN-CHIEF
Simon Lindgren

SENIOR EDITOR
Mattias Derlén

SENIOR EDITOR
Markus Naarttijärvi

SENIOR EDITOR
Mathilda Åkerlund

EDITOR
Karin Danielsson

EDITOR
Moa Eriksson Krutrök

EDITOR
Fateme Moradi

EDITOR
Christian Schwieter

EDITORIAL ASSISTANT
Emma Östin

JOURNAL METADATA

ISSN
2003-1998

DOI
<http://doi.org/10.33621/jdsr>

WEB
www.jdsr.io

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.

CONTACT US

E-MAIL
editor@jdsr.se

INTERNATIONAL EDITORIAL BOARD

Jean Burgess
QUEENSLAND UNIVERSITY OF TECHNOLOGY, AUSTRALIA

Mark Carrigan
UNIVERSITY OF CAMBRIDGE, UK

Nathalie Casemajor
INSTITUT NATIONAL DE LA RECHERCHE SCIENTIFIQUE, CANADA

Nick Couldry
LONDON SCHOOL OF ECONOMICS, UK

José van Dijck
UTRECHT UNIVERSITY, THE NETHERLANDS

Charles Ess
UNIVERSITY OF OSLO, NORWAY

Christian Fuchs
UNIVERSITY OF WESTMINSTER, UNITED KINGDOM

David Garcia
COMPLEXITY SCIENCE HUB VIENNA, AUSTRIA

David Gauntlett
RYERSON UNIVERSITY, TORONTO, CANADA

Tim Jordan
UNIVERSITY COLLEGE LONDON, UK

Anette Markham
AARHUS UNIVERSITY, DENMARK

Safiya Umoja Noble
UNIVERSITY OF SOUTHERN CALIFORNIA, USA

Sarah Pink
MONASH UNIVERSITY, AUSTRALIA

Thomas Poell
UNIVERSITEIT VAN AMSTERDAM, THE NETHERLANDS

Sarah T. Roberts
UNIVERSITY OF CALIFORNIA, LOS ANGELES, USA

Molly Wright Steenson
CARNEGIE MELLON SCHOOL OF DESIGN, PITTSBURGH, USA

Johanna Sumiala
UNIVERSITY OF HELSINKI, FINLAND

LICENCE & COPYRIGHT

JDSR is published under a Creative Commons BY-SA licence.
Cover photo by Dave Hoefler (Unsplash.com)

(Un)stable diffusions

The publics, publicities, and publicizations of generative AI

Fenwick McKelvey¹, Joanna Redden², Jonathan Roberge³, and Luke Stark²

¹ Concordia University, Canada

² Western University, Canada

³ Institut national de la recherche scientifique, Canada

✉ fenwick.mckelvey@concordia.ca

Abstract

Generative AI is a uniquely public technology. The large language models behind ChatGPT and other tools that generate text and images is a major develop in publicity as much as technology. Without public data and public participations, these large models could not be trained. Without the attention, hype, and hope around these technologies, the big AI firms probably could not afford the computational costs to train these models. Our special issue questions how Critical AI Studies can attend to the publics, publicities, and publicizations of generative AI. We situate AI's publicity as mode of publicity – hype, scandals, silences, and inevitability – as well as a mode of participation seen in the grown importance of technology demonstrations. Within this situation our contributions offer four different research paths: (1) situating the legacy media as an enduring process of legitimation; (2) looking at the ways that AI has a private life in public; (3) questioning the post-democratic future of public participation; and, (4) developing new prototypes of public participation through research creation.

Keywords: AI; generative AI; publicity; public theory; media and communication

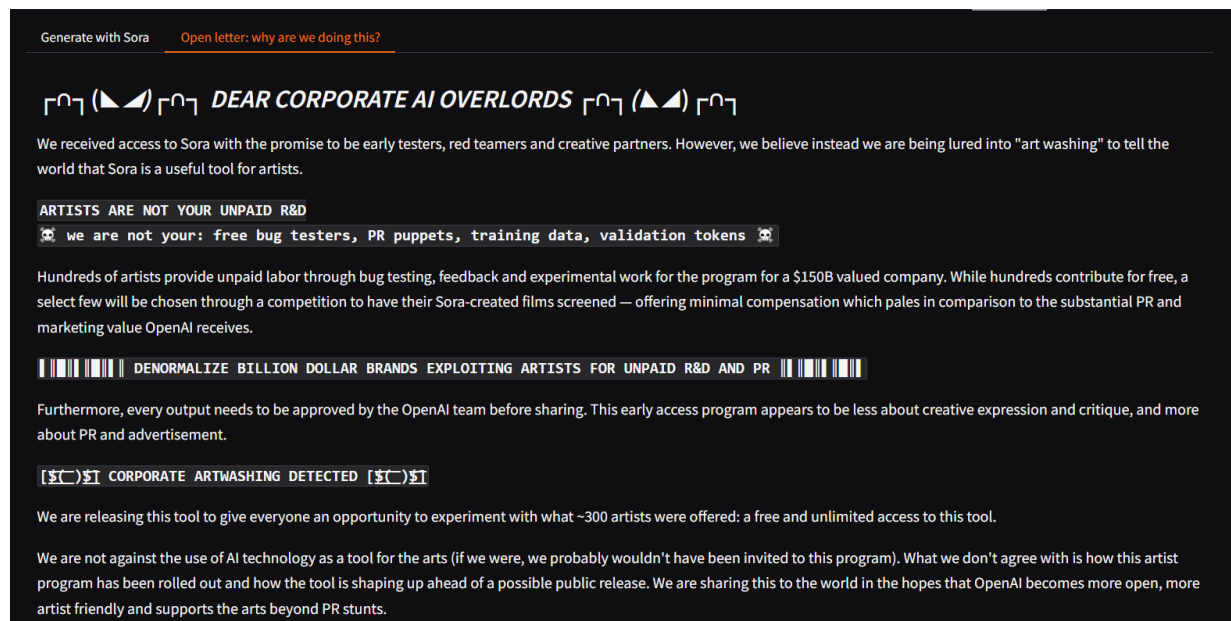


Figure 1. https://x.com/legit_rumors/status/1861431113408794898/photo/1 from <https://huggingface.co/spaces/PR-Puppets/PR-Puppet-Sora>

The revolt of the beta testers?

The message appeared on Hugging Face, a popular AI development site, in late November of 2024 along with a link to access OpenAI's unreleased model, Sora. The message was simultaneously a leak, a warning, and a conundrum. The post provided access to a new AI model in beta testing. The beta-testers, presumably artists, warned against "corporate artwashing" and undermining "creative expression." Laudable goals, but merely sharing a link that could be shut down easily (and quickly was)? Past hacktivism might have released unredacted documents, source code, or software, moving private or proprietary material into the public domain. What might have been the logic behind the leak?

This leak helps to introduce AI's unstable diffusions in the public sphere and our intent to encourage scholarship that disrupts the status quo approach to AI's social shaping (cf. Ananny, 2024; Gourlet et al., 2024; McQuillan, 2022). These upset beta-testers exploited OpenAI's carefully controlled spectacle of product launch "shipmas" in December (cf. Galloway & Thacker, 2007). OpenAI's marketing strategy exemplifies the problem Jodi Dean identified when warning against "the conviction that the solution to any problem is publicity" (Dean, 2001, p. 625). In releasing ChatGPT, OpenAI broke rank with a technocratic consensus around AI safety in favor of a public media blitz (For more behind its launch, see: Metz, 2023). OpenAI's success, like most tech firms, involves orchestrating product launches as strategic spectacles that give publics a chance to interact and play with new technologies. The novelty of playing with ChatGPT for the first time paid off handsomely, after all. Launches, however, are anything but a participatory public conversation. Instead, OpenAI accelerated technoscientific capitalism's fantasy of participation without power, whereby a product launch is merely a public invitation to occupy a future developed in a closed research lab (Dean, 2008; Jones & McKelvey, 2024; Palmås & Surber, 2022).

The revolt is a reminder of the ubiquity of communicative capitalism and the discursive power of publicity. If OpenAI exemplifies a certain mode of publicity, the revolt is a critical publication, a refusal to stay silent before yet another much-hyped launch. But in seeking publicity of their own, the revolt of the beta-testers also shows how important publicity is as a media tactic for resistance. The leak, if nothing else, shows that AI's publicity remains a point of contention and struggle.

Inspired by ChatGPT's launch and public ferment around generative AI more generally, our special issue turns to the concepts of publics, publication, and publicity as key concepts for the future of critical AI studies (Bode & Goodlad, 2023; Lindgren, 2023; Raley & Rhee, 2023; Roberge & Castelle, 2021). Our core motivation for this special issue has been to understand the social shaping of AI through its various modes of being brought to public attention (Barney, 2014; Marres, 2005). Historically, technologies have been important "material things" that serve as "crucial tools or props for the performance of public involvement in an issue" (Marres, 2010, p. 179). Generative AI today is precisely a shift in the "material and infrastructural conditions that give rise to publics and sustain their dynamics" (Ananny, 2024, p. 92).

What distinguishes AI studies now from pre-generative AI is mostly a matter of publicity. Generative AI is a uniquely public technology (McKelvey, 2022; McKelvey & Hunt, 2023). Technologies include OpenAI's GPT as well as numerous text-to-image generation software programs such as DALL-E and Stable Diffusion. All these tools have been released to the public for free use, or at least until their commercial equivalents are launched (Fingas, 2023). Three key traits unite GenAI:

1. a reliance on massive data sets largely collected from public data sources (e.g., the entire Internet) under dubious interpretations of fair dealing and fair use (Ferrari et al., 2023);
2. open feature sets that mimic common sense and human judgement (Natale, 2021); and
3. ease of access and use or public availability (Widder et al., 2024).

AI, in short, has been greenlit, and so it's here where we find ourselves: in public.

Generative AI, however, challenges our theories of publics and publicity. Historically, these relations have been mediated by the press (Schudson, 2008), but GenAI's arrival coincides with a destabilized media industry (Callison & Young, 2020; Curran, 2019), increasingly bypassed by mass-personal relations as in the case of ChatGPT's public launch (Bennett & Manheim, 2006; Gehl & Lawson, 2022). These relations are increasingly called into question as the scholarship reconciles with its overemphasis on active participation (e.g., posting or reacting) at the expense of more passive forms like listening, which are harder to measure (Ruess et al., 2021, p. 1507). Dean's warning seems ever more prescient:

The fantasy of abundance covers over the way facts and opinions, images and reactions circulate in a massive stream of content, losing their specificity and merging with and into the data flow. Any given message is thus a contribution to this ever-circulating content. (Dean, 2008, p. 58)

At a time when people are falling in love with chatbots (Rambukkana, 2021), the Internet has zombified into its own simulacra (Koebler, 2024), and there is a growing expectation to always be posting (Marwick, 2013), why does publicity even matter (cf. Couldry, 2022)?

We argue that publicity matters because the production of hype and its associated publicities *is* a strategy to maintain power and direct policy and political responses. As detailed by Clea Bourne, promotional culture is a central component of the AI value chain linking investment markets, consumer markets, and the state. This chain is facilitated by the

tech sector's capture of global knowledge apparatuses and modes of publication. This capture now incorporates control over digital media infrastructure and over promotional culture which funds that infrastructure, on into investment markets where, currently, tech companies dominate private equity investment portfolios as well as many of the world's major stock exchanges. (2024, p. 765)

Now media participation is more what Eeva Luhtakallio and Taina Meriluoto call a fame-based public sphere where the "quest for maximum visibility overrides the civic virtues of debating between equals and placing the collective before the individual" (2024, p. 335). Chasing the next status update seems precisely the kind of hype that Silicon Valley firms cultivate in their colonization of technological fantasies (Frase, 2016; Roberts & Hogan, 2019). Hype, however, is just one way to consider power and publicities. Our special issue gathers an incisive set of scholars to question AI's publicities.

Our introduction aims to situate the special issue's contributions as best as we can. We find inspiration from Mike Ananny who argues that generative AI is: “simultaneously (1) an object of public concern—a thing that demands collective oversight—and (2) an agent in public life—a semi-autonomous system with the power to mimic, engage, and structure the messages and rituals meant to care for collective life” (2024, p. 90). Our introduction takes on both sides, first looking at how AI has become, problematically, is an object of public concern and, second, at how AI's publicity offers specific scripts of public participation. These two default options situate our authors' contributions as ways of thinking through AI's many publicities.

Hype, scandal, inevitability, and silence: AI as an object of publicity

The discourse around AI is capacious enough to train its own large language model (for some examples, see: Bunz & Braghieri, 2022; Dandurand et al., 2023; Köstler & Ossewaarde, 2022; Nguyen & Hekman, 2022; Ulnicane et al., 2021). AI discourse, we argue, generally draws on four frames: hype, scandal, inevitability, or silence (Brennen, 2018, 2018; Sun et al., 2020), with the most prominent being hype (Ananny & Finn, 2020; Bourne, 2024; Roberge et al., 2020).

Hype is part of a long history of the technological sublime (Carey & Quirk, 1970; Mosco, 2004). AI builds on established ideas like the “fourth industrial revolution” and “disruption” that function as self-fulfilling prophecies. The economically fragile institution of journalism has largely fallen for this mode, due to the complexities of translating facts about these technologies to the public and the consolidation of newsrooms increasingly catering to business audiences (Dandurand et al., 2023). The consequence is that these technologies almost always arrive as good news in journalistic accounts. The launch of ChatGPT embodied the hype mode, fitting into a well-worn mode of a “normative framework of publicity [that is] drained of its critical value, and convert[ed] from a democratic asset to a democratic liability” (Barney, 2008, p. 92).

Hype, however, is not always positive. What STS scholar Lee Vinsel (2021) describes aptly as “criti-hype” pervades public discourse around generative AI. Past research on AI has shown how legacy media functions in the legitimization of technology futures, most evident in the turn toward AI doomsday theory in reaction to the 2023 Open Letter to Pause AI Experiments. The letter—a public demonstration of concerns—framed months of subsequent media coverage of AI (McKelvey, 2023). Criti-hype often lacks historicity, specifics, or critical analysis. We are of course not Pollyannas regarding the disruptive effects of generative AI tools. However, it is worthwhile to distinguish between forms and formats of AI criticism that inform, educate, and empower the public and those that obfuscate, handwave, or centre the speaker and not the problem at hand. Criti-hype has also enabled ouroboros-like circuits of claims and counter-claims regarding sensationalism, doomerism, and other prominent public affects around generative AI tools (Geburu & Torres, 2024; McKelvey & Roberge, 2023). Resistance to criti-hype does not mean acquiescence to the status quo. As Deleuze quipped, it's “not a question of worrying or of hoping for the best, but of finding new weapons” (1992, p. 4).

Scandals—or what we refer to as perceived proofs of social transgressions—are a pronounced feature of contemporary technological coverage (Lull & Hinerman, 1998; Trottier, 2017). Scandals are a driving force in modern technology governance (Bossetta, 2020). These moments of public outrage result from a mutually reinforcing relationship between newsrooms looking for easy, high-engagement stories and the affordances of social media, largely functioning as a distraction from other tasks (Blanchett et al., 2022; McKelvey et al., 2018). Scandals problematically depend on instinctive moral positions that accentuate Sun-Ha Hong's concept of techno-conservatism that projects the present into the future, and they do not necessarily involve opportunities for public engagement or democratic praxis (Hong, 2020). Without diminishing organized labour's success in raising AI as an issue in Hollywood, the reliance on machine learning as a theft of creative works is a scandalous frame that may eventually box in the movement or at least weaken its bargaining position to a matter of compensation for being included in training data sets.

The result of these fractured and anemic modes of publicity is a growing sense of *inevitability*. AI discourses are dominated by technology firms, government representatives, AI investors, global management consultancies, and think tanks. Like hype, inevitability advances a political project that nullifies public engagement as unnecessary or public opinion as out of touch (Jones & McKelvey, 2024; Palmås & Surber, 2022). These voices advocate faith in data-driven systems to address social problems while also increasing efficiency and productivity (Amoore et al., 2024; Fourcade & Gordon, 2020). Such discourses are used to reinforce the idea that the increasing use of AI applications across all spheres of life is inevitable, while sidelining or ignoring meaningful engagement with the ways these applications cause harm (Walker 2022).

If not positive, AI coverage is marked by *silences* or, in effect, closures when aspects of AI remain too uncontroversial to report (Suchman, 2023). Some modes of publicity negate public attention, so-called cold controversies. (Callon, 1998; Venturini et al., 2015). Silences are a result of AI's cold controversiality where a presumed consensus prevents the publication of dissent and debate (Dandurand et al., 2023). National imaginaries of AI that presume the technologies' geopolitical and economic importance legitimate these silences becoming an unspoken inevitability (Bareis & Katzenbach, 2021).

These default modes of publicities undermine public agency and democracy, cultivating a permissive fear. The narrative of fear led to a stock market spike, with the "AI boom" being credited as fuelling the best first quarter world stocks had seen in five years (Steer et al. 2024). The investment boom was driven by industry players like Morgan Stanley positioning AI as having a "breakthrough moment" and stressing that investors had an opportunity to tap into a \$6 trillion opportunity (Morgan Stanley, 2023). While the generative hype bubble may be deflating in the midst of talk that it may have been overvalued, the harmful effects of AI continue (Widder & Hicks, 2024). As argued by David Widder and Mar Hicks, we can already see the harmful effects of the latest AI information cycle. This can be seen in wages and salaries stagnating or decreasing in response to threats of being replaced by AI. Countries around the world are building new fossil fuel-based power plants and delaying the phasing out of coal to meet the anticipated needs of AI (Volcovici et al., 2024). Researchers are detailing the implications of GenAI for our information ecosystems, raising concerns about how content pollution will affect people and society (Mitra et al., 2024). Awareness of these concerns and many of the limits of generative AI pre-date the launch of Chat GPT (Bender et al., 2021), leading scholars to wonder:

[W]hat if the hype was always meant to fail? What if the point was to hype things up, get in, make a profit, and entrench infrastructure dependencies before critique, or reality, had a chance to catch up? (Widder and Hicks 2024)

This proposition builds on recognition of the multifaceted ways that AI hype is embedded within tech sector business models and venture capital strategies and facilitated through global communicative infrastructure. To consider AI ethics, argues Bourne (2024), requires examining promotional ethics and critical investigation of "the processes used to drive hype cycles" so that we might develop better strategies to resist them (2024, p. 765).

These modes of publicity emphasize the legacy media's role co-shaping AI's attending discourse, but what about the technology itself? How has GenAI been made available to the public? AI's publicity also involves how the technology is enrolled in acts of public participation like the spectacular failure of Microsoft's Tay AI (Neff & Nagy, 2016). These mistakes might diminish the power of firms to produce mass public technological demonstrations (Halpern et al., 2017; Rosental, 2021). Like Steve Jobs commanding public attention by pulling a laptop out of a manila envelope, now AI model-makers release new tools like Sora in the hopes of capturing users' attention and public interest. Returning to Ananny's argument that AI is also an agent in public life, we wish to emphasize the consequences of AI directly addressing (or one-step flow relations) publics. Far from an organic relationship, we wish to highlight how technology and technology firms engineer AI's publicities today.

Engineering publics: Demonstration and direct public relations

Demonstrations are a novel thematic to explore ongoing issues of democratic legitimacy and backsliding associated with technological governance (Halpern & Mitchell, 2022; Rosental, 2021). A demonstration is both “reality and a fiction” that through “emotional work” aims to “control the audience interpretations” (Rosental, 2021, p. 17). Technological firms depend on demonstrations as critical public interventions that shape a global politics of expectations and futures-making. AI functions as what Morgan Ames calls a charismatic technology that “promises to transform its users' sociotechnical experience” (Ames, 2019, p. 11). Even talking about generative AI seems caught in its own charismatic future, and its precisely these kinds of engineered public relations that allow for publics to access AI that we wish to call attention to, the demonstration.

AI is marked by a faith in public demonstration. Consider the unicorn AI firm of the European Union, Mistral. To test its latest large language model, Mistral AI released it to the public, free and uncensored, just as all G7 members ranked disinformation as the top threat posed by GenAI. The firm, one of Europe's top AI start-ups, explained that the model “does not have any moderation mechanism. We're looking forward to engaging with the community on ways to make the model respect guardrails, allowing for deployment in environments requiring moderated outputs” (Maiberg, 2023). Mistral's “open source” approach, shared by many in the AI industry, is a profound shift in building trustworthy technology, indeed a shift in the production of trust itself.

Demonstrations have their own social power. Generative AI, in less than eight months, has prompted a wholesale re-evaluation of labour, education, and government, requiring rapid revision of AI legislation years in the making. Even OpenAI CEO Sam Altman admitted the risks of relying on demos to market AI. In a recent keynote, Altman explains: “One of the strange things about these technologies is that they are impressive but not robust . . . so you use them in the first demo; you have this [impressive reaction]. But [after a hundred times], you see the weaknesses” (Kahn, 2023). His admission encapsulates another of our issue's goals: to critically investigate the power of the demonstration as a matter of governance, media systems, and trust. Despite the global influence of demonstrations and charismatic technologies, traditional civic epistemologies struggle to counteract the bias and political economic pressures behind them. From OpenAI's ChatGPT to Apple's latest VR, large technology firms and consultancy firms are adept at demos and demoing that steers collective future-making whereas democratic institutions and publics increasingly lack similar capacities to demonstrate plausible, desirable futures.

New directions in critical studies of AI's publicities

Our special issue collects new directions in the study of AI's publicities. Across these pieces, the authors have made suggestions for various interventions or reflections on critical practices, such as community engagement and mobilization, futures literacy, or capacity building, in the cause of decentering the strategic futures-making by large technology firms. Our call asked for three broad themes:

- Critical and comparative studies of AI's publicities with regards to the launch and hype of AI. We particularly welcome papers that focus on cases outside the Global North;
- Ethnographic, discursive, or engaged research with AI's public such as AutoGPT, HustleGPT, or other publics forming around the use, misuse or resistance to GPAI;
- Interventions or reflections on critical practices, such as community engagement and mobilization, futures literacy, or capacity building, for better publicizations of AI de-centering the strategic futuring employed by large technology firms.

From these broad openings, contributions clustered around four key themes:

1. situating the legacy media as an enduring process of legitimation;
2. looking at the ways that AI has a private life in public;
3. questioning the post-democratic future of public participation; and,
4. developing new prototypes of public participation through research creation.

These themes and their contributions offer novel approaches to advance the relationship between AI and public theory in this special issue and hopefully for special issues to come.

First, AI's publicities can be attended to through a renewed interest in legacy media's legitimation function. Legitimation, in turn, concerns how coverage frames the situation, in effect, cocreating affected subjectivities as stories and emotion. Is an individual's perception of ChatGPT's outputs an exception to the norm? Journalism has an important function as a reflexive apparatus in the legitimation of publics and affects. The reflexive apparatus is increasingly important in technologically affected publics isolated by time and space from their peers (McKelvey, 2014, 2018). Turmoil in media systems offers a renewed opportunity to consider globally the place of the press in the controversies (or lack thereof) around generative AI. At first glance, generative AI has marked a moment of growing acknowledgement of AI's risks.

Lensa, the subject of Kate Miltner's contribution, offers a good example of how to study AI's publicities through thematic media analysis. Lensa was one of the earliest apps to integrate generative AI image functions, allowing its users to create new profile pictures. Given the biases in its underlying model (Stable Diffusion), the app was an important moment in how the press legitimated the model's functionalities as either structural biases or accidental hallucinations. Miltner finds that coverage presented the model's harms as another example of "just the way things are," nurturing a culture of political resignation around these technologies' social harms. These findings are not surprising; rather, they contribute to what Callison and Young (2020) call journalism's reckoning, where the field must better account for its functions and limits and, in this case, its responsibilities to its publics.

Studying the media's legitimation function for AI extends beyond AI as a tool. The *New York Times*, as well as most major news outlets in Canada, has broken with its peers and sued OpenAI rather than licensing its content. The legal case best encapsulates the existential challenge generative AI presents to journalism. More structurally, if journalism has been a tool to produce audiences for advertisers, then the ad-supported ecosystem seems largely indifferent to the content of the news, AI-generated or not (Braun & Eklund, 2019). Court cases punctuate the need to restate the value of news as something other than content for AI engines. Here the press, along with other creative industries, becomes one of the few places capable of contesting the economic narrative of AI as the foundation of a beneficent industrial revolution.

A second facet of AI's publicities attends to AI's mass-personal publicities beyond the demonstration. Noortje Marres (2010) highlighted how individual encounters with technology could become the basis for public awareness—an energy efficiency rating on an appliance becomes a potential link to an environmental movement. Without overstating that possibility, a second avenue considers how AI technology actively participates in its own sense making. Mike Ananny writes, "GenAI's distinct public quality comes from its capacity to participate in shaping the very conversations that might hold it accountable" (Ananny, 2024, p. 90). AI's hyperproduction becomes the possibility for its outputs to exceed the parameters set in a demonstration mode (Ferrari & McKelvey, 2022). Much of the work of firms like OpenAI is to manage such generative hyperproduction, channeling its excessive outputs into plausible corporate futures, or what Sun-Ha Hong (2021) calls a techno future of conservatism that preserves the present order into the future. Can we look, following Ananny (2023), for failures or other ways to think about how AI becomes public through less intentional means than the demo?

These moments have grown along with a demonstration society that has largely let publics experiment with their own limited AI futures. The challenge of the demo, as discussed above, remains the ways that

its procedural rhetorics and charismatic functions frame the technological encounter. Etienne Grenier and Nicolas Chartier-Edwards in their contribution delve into perhaps the darkest of AI un-publicities, AI-powered hiring practices, to consider how a deeper political economy primes the early encounter with AI. Their attention to corporate storytelling and user experience helps render new publicities sought to be outdone by public relations. Scholarship here returns to Stuart Hall's "conjuncture" to describe the new public formations in process. Fabian Offert, by contrast, takes a more considered study of racialized encounters with Stable Diffusion's feature space, when immediate personal discomfort becomes a bug to be fixed rather than a moment to reckon with the "whiteness" of visual culture itself.

Building on feminist standpoint theory's emphasis on the sites of marginalization, future studies could show how labour, particularly precarious labour, is an important moment in AI's publicities. These moments when AI is one's boss or coworker form a public encounter that has been cultivated to arrive as normal on arrival. Such an approach rejects what easily could be another largely consumer-oriented focus on AI's impacts—perhaps a legacy of past research into network neutrality that let consumer entitlement stand in for calls for a right to access a technology. Publics do form around these moments—like "Batterygate" when researchers and publics discovered that Apple slowed down older phones (Shi, 2018). These moments can almost feel nostalgic under the modern YouTube hustle aesthetic, where learning is part of a pyramid scheme to unlock passive income. Labour is perhaps a more productive site of study than questioning the publics tagged in the next #GPTHustle.

A third approach to AI's publicities could make the link more directly and question AI's role in post-democratic public participation. How does the rhetoric of participation indicate the ubiquity of information feudalism and a new courtly mode of politics for this new AI-generated spirit of laws? Our special issue is an artefact of a moment of uncertainty about publicity more broadly. Traditionally, media systems have been considered crucial to public participation, and yet artificial intelligence compounds an ongoing legitimacy crisis. Amidst widespread democratic backsliding, increasingly global success for right wing cryptopolitics, and the popularity of right-wing AI death cults, perhaps the time to look for the public sphere has passed. In its stead, we can turn toward how the speculation functions of generative AI enable new modes of dark participation—a kind of inverse Roko's basilisk where we look at how current society has oriented itself around the coming of an AI superintelligence. Though only adjacent to political communication, the problem of post-democratic public participation involves a similar suspension of traditional norms in publics theory to comprehend the productive fantasies of generative AI when "democracy itself" has become "a contested value" (Knüpfer et al., 2024, p. 1011).

Many of the articles in our special issue take up AI's publicities as public drama. In their contribution, Xerxes Minocher considers the drama in and among Go players in reaction to AI "solving" the famously complex game. The drama, which happened largely outside the mainstream press, entailed critical imagination of AI's use and a moment of reconstruction. Alison Powell and Fenwick McKelvey reflect on doing critical policy work now. Where the field of critical policy studies is turning to AI (Paul, 2022), at a time of widespread consultation theatre. Their efforts cut against the policy sciences tradition of evidence-based decision making to instead question the preconstituted roles to be played in the drama of AI governance. Sun-Ha Hong's contribution similarly considers the drama of GenAI that advances our earlier discussion of demonstrations and the engineered public relations. What Hong abduction "serves strategic objectives for shaping the terms of the AI conversation in favour of the incumbent tech elites," and, in so doing, calling into questions the assumptions about public debate as a means to democratic participation. Finally, Gwendolyn Blue and Mél Hogan interrogate how scientists act out their responsibilities through declarations of principles that maintain the self-regulation of AI by letting the symbolic role of principles ensure that only experts are empowered to participate. Drama, in its many forms, offers a third facet to investigate AI publicities by reformulating what counts as public participation.

A final mode of publicity draws on the research creation tradition to experiment with new prototypes for public participation with and through generative AI. Our contributions join growing calls for a

renewed strategies of participation in an age of AI (Gourlet et al., 2024; Marres, 2024). Crystal Chokshi contributes exactly such a piece through the creation of an interface that presents its users with the climate impacts of generative AI. Here we can end on a more hopeful note. With ample evidence of the limitations of AI's publicities, how might scholarship endure as a site of experimentation, indeed, as a demonstration of the importance of the public to societies invested in rendering them as merely a surplus population?

These articles advance how critical AI studies must address the publicity inherent in generative AI. The challenge is two-fold, then, to not only diagnose the current modes of AI's publicities but to consider alternative and normative futures for our public media systems. To get *there*, however, we must first attend to AI today. These efforts are the beginning of a renewal of public theory around generative AI and the broader democratic downturn, a small turn in the opposite direction of this downward spiral undoing the demos (Brown, 2015).

References

- Ames, M. G. (2019). *The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child*. MIT Press.
- Amoore, L., Campolo, A., Jacobsen, B., & Rella, L. (2024). A world model: On the political logics of generative AI. *Political Geography*, 113, 103134. <https://doi.org/10.1016/j.polgeo.2024.103134>
- Ananny, M. (2023). Making Mistakes. *Osiris*. <https://doi.org/10.1086/725146>
- Ananny, M. (2024). Making Generative Artificial Intelligence a Public Problem. Seeing Publics and Sociotechnical Problem-Making in Three Scenes of AI Failure. *Javnost - The Public*, 31(1), 89–105. <https://doi.org/10.1080/13183222.2024.2319000>
- Ananny, M., & Finn, M. (2020). Anticipatory news infrastructures: Seeing journalism's expectations of future publics in its sociotechnical systems. *New Media & Society*, 22(9), 1600–1618. <https://doi.org/10.1177/1461444820914873>
- Bareis, J., & Katzenbach, C. (2021). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values*, 01622439211030007. <https://doi.org/10.1177/01622439211030007>
- Barney, D. (2008). Politics and Emerging Media: The Revenge of Publicity. *Global Media Journal - Canadian Edition*, 1(1), 89–106.
- Barney, D. (2014). Publics without Politics: Surplus Publicity as Depoliticization. In K. Kozolanka (Ed.), *Publicity and the Canadian State: Critical Communications Approaches* (pp. 72–88). University of Toronto Press.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bennett, W. L., & Manheim, J. B. (2006). The One-Step Flow of Communication. *The Annals of the American Academy of Political and Social Science*, 608(1), 213–232.
- Blanchett, N., McKelvey, F., & Brin, C. (2022). Algorithms, platforms, and policy: The changing face of Canadian news distribution. In J. Meese & S. Bannerman (Eds.), *The Algorithmic Distribution of News*. Palgrave Macmillan. <http://link.springer.com/book/9783030870850>
- Bode, K., & Goodlad, L. M. E. (2023). Data Worlds: An Introduction. *Critical AI*, 1(1–2). <https://doi.org/10.1215/2834703X-10734026>
- Bossetta, M. (2020). Scandalous Design: How Social Media Platforms' Responses to Scandal Impacts Campaigns and Elections. *Social Media + Society*, 6(2), 205630512092477. <https://doi.org/10.1177/2056305120924777>
- Bourne, C. (2024). AI hype, promotional culture, and affective capitalism. *AI and Ethics*, 4(3), 757–769. <https://doi.org/10.1007/s43681-024-00483-w>
- Braun, J. A., & Eklund, J. L. (2019). Fake News, Real Money: Ad Tech Platforms, Profit-Driven Hoaxes, and the Business of Journalism. *Digital Journalism*, 0(0), 1–21. <https://doi.org/10.1080/21670811.2018.1556314>
- Brown, W. (2015). *Undoing the Demos: Neoliberalism's Stealth Revolution*. Zone Books.
- Bunz, M., & Braghieri, M. (2022). The AI doctor will see you now: Assessing the framing of AI in news coverage. *AI & Society*, 37, 9–22. <https://doi.org/10.1007/s00146-021-01145-9>
- Callison, C., & Young, M. (2020). *Reckoning: Journalism's limits and possibilities*. Oxford University Press.
- Callon, M. (1998). An essay on framing and overflowing: Economic externalities revisited by sociology. *The Sociological Review*, 46(S1), 244–269. <https://doi.org/10.1111/j.1467-954X.1998.tb03477.x>
- Carey, J. W., & Quirk, J. J. (1970). The Mythos of the Electronic Revolution. *The American Scholar*, 39(2), 219–241. JSTOR.
- Couldry, N. (2022). Post-Covid: What is cultural theory useful for? *International Journal of Cultural Studies*, 25(3–4), 253–259. <https://doi.org/10.1177/13678779211055846>
- Curran, J. (2019). Triple crisis of journalism. *Journalism*, 20(1), 190–193. <https://doi.org/10.1177/1464884918807034>

- Dandurand, G., McKelvey, F., & Roberge, J. (2023). Freezing out: Legacy media's shaping of AI as a cold controversy. *Big Data & Society*, 10(2), 20539517231219242. <https://doi.org/10.1177/20539517231219242>
- Dean, J. (2001). Publicity's Secret. *Political Theory*, 29(5), 624–650.
- Dean, J. (2008). *Communicative Capitalism: Circulation and the Foreclosure of Politics* (M. Boler, Ed.; pp. 101–122). MIT Press.
- Deleuze, G. (1992). Postscript on the Societies of Control. *October*, 59(1), 3–7.
- Ferrari, F., & McKelvey, F. (2022). Hyperproduction: A social theory of deep generative models. *Distinktion: Journal of Social Theory*, 0(0), 1–23. <https://doi.org/10.1080/1600910X.2022.2137546>
- Ferrari, F., van Dijck, J., & van den Bosch, A. (2023). Foundation models and the privatization of public knowledge. *Nature Machine Intelligence*, 5(8), 818–820. <https://doi.org/10.1038/s42256-023-00695-5>
- Fingas, J. (2023, January 11). *OpenAI will soon test a paid version of its hit ChatGPT bot*. <https://finance.yahoo.com/news/openai-chatgpt-professional-paid-chatbot-143004442.html>
- Fourcade, M., & Gordon, J. (2020). Learning Like a State: Statecraft in the Digital Age. *Journal of Law and Political Economy*, 1(1). <https://doi.org/10.5070/LP61150258>
- Frase, P. (2016). *Four Futures: Life After Capitalism*. Verso.
- Galloway, A., & Thacker, E. (2007). *The Exploit: A Theory of Networks*. University of Minnesota Press.
- Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*. <https://doi.org/10.5210/fm.v29i4.13636>
- Gehl, R. W., & Lawson, S. T. (2022). *Social Engineering: How Crowdmasters, Phreaks, Hackers, and Trolls Created a New Form of Manipulative Communication*. MIT Press.
- Gourlet, P., Ricci, D., & Crépel, M. (2024). Reclaiming artificial intelligence accounts: A plea for a participatory turn in artificial intelligence inquiries. *Big Data & Society*, 11(2), 20539517241248093. <https://doi.org/10.1177/20539517241248093>
- Halpern, O., Mitchell, R., & Geoghegan, B. D. (2017). The Smartness Mandate: Notes toward a Critique. *Grey Room*, 68, 106–129. https://doi.org/10.1162/GREY_a_00221
- Hong, S. (2020). *Technologies of speculation: The limits of knowledge in a data-driven society*. New York University Press.
- Hong, S. (2021). Technofutures in Stasis: Smart Machines, Ubiquitous Computing, and the Future That Keeps Coming Back. *International Journal of Communication*, 15(0), Article 0.
- Jones, M., & McKelvey, F. (2024). Deconstructing public participation in the governance of facial recognition technologies in Canada. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-024-01952-w>
- Kahn, J. (2023, January 25). *The inside story of ChatGPT: How OpenAI founder Sam Altman built the world's hottest technology with billions from Microsoft*. *Fortune*. <https://fortune.com/longform/chatgpt-openai-sam-altman-microsoft/>
- Knüpfer, C., Jackson, S. J., & Kreiss, D. (2024). Political Communication Research is Unprepared for the Far Right. *Political Communication*, 41(6), 1009–1016. <https://doi.org/10.1080/10584609.2024.2414268>
- Koebler, J. (2024, May 2). *Facebook's AI Spam Isn't the 'Dead Internet': It's the Zombie Internet*. 404 Media. <https://www.404media.co/facebook-ai-spam-isnt-the-dead-internet-its-the-zombie-internet/>
- Köstler, L., & Ossewaarde, R. (2022). The making of AI society: AI futures frames in German political and media discourses. *AI & SOCIETY*, 37(1), 249–263. <https://doi.org/10.1007/s00146-021-01161-9>
- Lindgren, S. (Ed.). (2023). *Handbook of critical studies of artificial intelligence*. Edward Elgar Publishing.
- Luhtakallio, E., & Meriluoto, T. (2024). Fame democracy? Social media and visibility-based transformation of the public sphere. *Distinktion: Journal of Social Theory*, 25(3), 318–342. <https://doi.org/10.1080/1600910X.2023.2263818>
- Lull, J., & Hinerman, S. (Eds.). (1998). *Media Scandals*. Columbia University Press.
- Maiberg, E. (2023, September 29). *\$260 Million AI Company Releases Undeletable Chatbot That Gives Detailed Instructions on Murder, Ethnic Cleansing*. 404 Media. <https://www.404media.co/260-million-ai-company-releases-chatbot-that-gives-detailed-instructions-on-murder-ethnic-cleansing/>
- Marres, N. (2005). *Issues Spark a Public into Being: A Key But Often Forgotten Point of the Lippmann-Dewey Debate* (B. Latour & P. Weibel, Eds.; pp. 208–217). MIT Press.
- Marres, N. (2010). Front-staging Nonhumans: Publicity as a Constraint on the Political Activity of Things. In B. Braun & S. J. Whatmore (Eds.), *Political Matter: Technoscience, Democracy, and Public Life* (pp. 177–210). University of Minnesota Press.
- Marres, N. (2024). Do automated vehicle trials test society? Testing mobility futures in the West Midlands. *Mobilities*, 0(0), 1–21. <https://doi.org/10.1080/17450101.2024.2356566>
- Marwick, A. E. (2013). *Status Update: Celebrity, Publicity, and Branding in the Social Media Age*. Yale University Press.
- McKelvey, F. (2014). Algorithmic Media Need Democratic Methods: Why Publics Matter. *Canadian Journal of Communication*, 39(4), 597–613.
- McKelvey, F. (2018). *Internet daemons: Digital communications possessed*. University of Minnesota Press.
- McKelvey, F. (2022, October 27). *<nettime> Thoughts on Stable Diffusion and Free Culture*. <https://www.nettime.org/Lists-Archives/nettime-l-2210/msg00025.html>

- McKelvey, F. (2023, April 3). *Let's base AI debates on reality, not extreme fears about the future*. The Conversation. <http://theconversation.com/lets-base-ai-debates-on-reality-not-extreme-fears-about-the-future-203030>
- McKelvey, F., Côté, M., & Raynauld, V. (2018). Scandals and Screenshots: Social Media Elites in Canadian Politics. In A. Marland, T. Giasson, & A. Lawlor (Eds.), *Political Elites in Canada: Power and Influence in Instantaneous Times* (pp. 204–222). UBC Press.
- McKelvey, F., & Hunt, R. (2023, May 19). *ChatGPT's Infrastructural Ambitions: AI, Commodification, and the Commons*. Centre for Media, Technology and Democracy. <https://www.mediatechdemocracy.com/all-work/chatgpts-infrastructural-ambitions-ai-commodification-and-the-commons>
- McKelvey, F., & Roberge, J. (2023). Recursive Power. In S. Lindgren (Ed.), *Handbook of critical studies of artificial intelligence* (pp. 21–32). Edward Elgar Publishing.
- McQuillan, D. (2022). *Resisting AI: An Anti-fascist Approach to Artificial Intelligence* (1st ed.). Bristol University Press. <https://doi.org/10.2307/j.ctv2rcnp21>
- Metz, C. (2023, March 31). The ChatGPT King Isn't Worried, but He Knows You Might Be. *The New York Times*. <https://www.nytimes.com/2023/03/31/technology/sam-altman-open-ai-chatgpt.html>
- Mitra, B., Cramer, H., & Gurevich, O. (2024). *Sociotechnical Implications of Generative Artificial Intelligence for Information Access* (No. arXiv:2405.11612; Version 1). arXiv. <https://doi.org/10.48550/arXiv.2405.11612>
- Morgan Stanley. (2023, April 18). *The \$6 Trillion Opportunity in AI*. Morgan Stanley. <https://www.morganstanley.com/ideas/generative-ai-growth-opportunity>
- Mosco, V. (2004). *The Digital Sublime: Myth, Power, and Cyberspace*. MIT Press.
- Natale, S. L. in C. and M. S. S. (2021). *Deceitful Media: Artificial Intelligence and Social Life After the Turing Test*. Oxford University Press.
- Neff, G., & Nagy, P. (2016). Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10, 4915–4931.
- Nguyen, D., & Hekman, E. (2022). The news framing of artificial intelligence: A critical exploration of how media discourses make sense of automation. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-022-01511-1>
- Palmãs, K., & Surber, N. (2022). Legitimation crisis in contemporary technoscientific capitalism. *Journal of Cultural Economy*, 15(3), 373–379. <https://doi.org/10.1080/17530350.2022.2065331>
- Paul, R. (2022). Can critical policy studies outsmart AI? Research agenda on artificial intelligence technologies and public policy. *Critical Policy Studies*, 16(4), 497–509. <https://doi.org/10.1080/19460171.2022.2123018>
- Raley, R., & Rhee, J. (2023). Critical AI: A Field in Formation. *American Literature*, 95(2), 185–204. <https://doi.org/10.1215/00029831-10575021>
- Rambukkana, N. (Ed.). (2021). *Intersectional automations: Robotics, AI, algorithms, and equity*. Lexington Books.
- Roberge, J., & Castelle, M. (Eds.). (2021). *The cultural life of machine learning: An incursion into critical AI studies*. Palgrave Macmillan. <https://doi.org/10.1007/978-3-030-56286-1>
- Roberge, J., Senneville, M., & Morin, K. (2020). How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data & Society*, 7(1), 205395172091996. <https://doi.org/10.1177/2053951720919968>
- Roberts, S. T., & Hogan, M. (2019). Left Behind: Futurist Fetishists, Prepping and the Abandonment of Earth. *B2o: An Online Journal*, 4(2). <https://escholarship.org/uc/item/8sr8n99w>
- Rosental, C. (2021). *The demonstration society* (C. Porter, Trans.). The MIT Press.
- Ruess, C., Hoffmann, C. P., Boulianne, S., & Heger, K. (2021). Online political participation: The evolution of a concept. *Information, Communication & Society*, 26(8), 1495–1512. <https://doi.org/10.1080/1369118X.2021.2013919>
- Schudson, M. (2008). *Why Democracies Need an Unlovable Press*. Polity.
- Shi, A. (2018, September 8). *Batterygate: A Complete History of Apple Throttling iPhones*. iFixit. <https://www.ifixit.com/News/11208/batterygate-timeline>
- Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, 10(2), 20539517231206794. <https://doi.org/10.1177/20539517231206794>
- Trottier, D. (2017). Scandal mining: Political nobodies and remediated visibility. *Media, Culture & Society*, 0163443717734408.
- Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W.-G. (2021). Framing governance for a contested emerging technology: insights from AI policy. *Policy and Society*, 40(2), 158–177. <https://doi.org/10.1080/14494035.2020.1855800>
- Venturini, T., Ricci, D., Mauri, M., Kimbell, L., & Meunier, A. (2015). Designing Controversies and Their Publics. *Design Issues*, 31(3), 74–87. https://doi.org/10.1162/DESI_a_00340
- Vinsel, L. (2021, February 1). You're Doing It Wrong: Notes on Criticism and Technology Hype. *Medium*. <https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5>
- Volcovici, V., Kearney, L., & Volcovici, V. (2024, November 26). Data-center reliance on fossil fuels may delay clean-energy transition. *Reuters*. <https://www.reuters.com/technology/artificial-intelligence/how-ai-cloud-computing-may-delay-transition-clean-energy-2024-11-21/>
- Widder, D. G., & Hicks, M. (2024). *Watching the Generative AI Hype Bubble Deflate* (No. arXiv:2408.08778). arXiv. <https://doi.org/10.48550/arXiv.2408.08778>

Widder, D. G., Whittaker, M., & West, S. M. (2024). Why ‘open’ AI systems are actually closed, and why this matters. *Nature*, 635(8040), 827–833. <https://doi.org/10.1038/s41586-024-08141-1>

“A.I. is holding a mirror to our society”

Lensa and the discourse of visual generative AI

Kate M. Miltner

University of Sheffield, United Kingdom

✉ k.miltner@sheffield.ac.uk

Abstract

Since the release of the open-source AI model Stable Diffusion in August 2022, a panoply of apps that create AI-generated portraits, or avatars, have exploded in popularity. One of the most notable examples is AI-powered photo editing app Lensa, owned by Prisma Labs. Lensa launched in 2018 but went viral in late 2022 due to the launch of its Magic Avatar feature, which uses Stable Diffusion to create fantastical (and occasionally bizarre) portraits of users. This paper analyzed the global English-language press coverage of Lensa and found that it focused on the app’s predatory data practices, the biased content it produced, and the user behaviors associated with it. I argue that this coverage provides evidence of discursive closure around key issues associated with visual generative AI that supports the maintenance of the status quo. I suggest that the press coverage of Lensa, which both articulates key AI-related harms and frames those harms as intractable and insolvable, creates a discourse of inevitability that has implications for how these issues are understood by the public, and for the approaches that are taken to address them. In doing so, it not only offers distinct advantages to those who stand to benefit most from this discourse, but also forecloses more imaginative public discussions of what visual generative AI could– or should– be.

Keywords: generative AI; discursive closure; selfies; algorithmic bias

1. Introduction

“Beware a world where artists are replaced by robots. It’s starting now” intoned a Los Angeles Times headline (Crabapple, 2022). “The inherent misogyny of AI portraits – Amelia Earhart rendered naked on a bed”, proclaimed The Guardian (Demopolous, 2022). “Your selfies are helping AI learn. You did not consent to this”, cautions The Washington Post (Ovide, 2022). All three of these headlines offer different warnings about the dangers of AI, and all three headlines are about the same app: Lensa, a photo editing app that makes AI-powered selfies.

Since the release of the open-source AI model Stable Diffusion in August 2022, a panoply of apps that create AI-generated portraits, or avatars, have exploded in popularity. Lensa, one of the most notable examples of this app category, originally launched in 2018 but went viral in late 2022 due to the addition of its Magic Avatar feature. Lensa uses the Stable Diffusion AI model to create fantastical (and occasionally bizarre) portraits of users, which are created by uploading a series of photographs and paying

a small fee; the app then generates anywhere from 50 to 200 AI-generated images in a series of user-selected themes, such as “Rock Star”, “Fairy Princess”, “Superhero”, and “Astronaut” (see Figure 1). After the Magic Avatar launch in November 2022, Lensa was installed over 13.5 million times worldwide, with users spending over USD\$29 million on the app in 12 days (McCluskey, 2022). In a matter of weeks, Lensa was the number one app in the Apple App Store’s Photo & Video charts, beating out heavyweights like YouTube and Instagram (Silberling, 2022). Lensa’s meteoric rise in popularity can be largely attributed to the fact that many users—including some public figures—posted their Magic Avatars on social media, sparking curiosity and a flood of global press coverage that both hyped and critiqued the phenomenon.



Figure 1. Examples of the author’s Magic Avatars

Despite the ostensible novelty of Magic Avatars, the press coverage about Lensa recycles common concerns about emerging technologies, particularly around algorithmic bias, unethical data use, and the narcissism of digital self-representation. While some of Lensa’s press coverage simply described the app and how to use it, a significant number of articles were highly critical, suggesting that the app engaged in predatory data practices, produced biased content, and encouraged user behaviors that were frivolous at best and harmful at worst. However, while the discourse about Lensa’s Magic Avatars articulates key AI-related harms, it also treats those harms as unfortunate but intractable problems that are already too difficult to solve. This suggests that some discursive closure (Deetz, 1992; Leonardi & Jackson, 2003; Markham, 2021) has already developed around both the risks and the potential of visual generative AI, which has implications for how these technologies are understood by the public. By relying on well-worn understandings of the key concerns raised by AI and other algorithmic technologies, the press coverage of Lensa creates a “discourse of inevitability” (Leonardi & Jackson, 2003; Markham, 2021) that makes these harms seem inescapable and unaddressable.

While Lensa’s moment was seen to be “over” (Lovejoy, 2023) in early 2023—indeed, the AI avatar app trend was said to have had “fizzled” (Perez, 2023) after a few months—these discourses remain as regular features in the press as competitor apps continue to launch and their features are hyped and occasionally dissected (e.g., Brown, 2023; Khan, 2023). As such, analyzing the discourse around Lensa offers an opportunity to reflect on the broader processes of AI publicization and their attendant implications for “exactly what AI is supposed to represent in the world” (Broussard et al., 2019). By promoting a discourse of inevitability around the harms of visual generative AI, the press coverage of Lensa offers “a particular view of reality that is maintained at the expense of equally plausible ones” (Deetz, 1992, p. 188); this ends up reinforcing narrow understandings of the issues at hand. In doing so, it not only contributes to the calcification of the broader discourse around AI and its “myths” (Natale and Ballatore, 2020) but also forecloses more imaginative public discussions of what visual generative AI could— or should— be.

2. The hype and harms of visual generative AI

Generative AI is a colloquial term for deep-learning computer models that can take a collection of data as an input (e.g., an artist's full body of work) and analyze that data to generate statistically probable outputs when provided with a prompt; these models "encode a simplified representation of their training data and draw from it to create a new work that's similar, but not identical, to the original data" (Martineau, 2021). In other words, generative AI models "find patterns in the data they are trained on and then create new work by attempting to mimic those patterns" (Miltner and Highfield, 2024).

There are a variety of generative AI models that can generate and/or modify a wide variety of media, including text, images, speech, video, and audio. Large language models (LLMs) – such as ChatGPT—produce text, and diffusion models are most often used in audiovisual modification and generation. Some of the most well known image-generating diffusion models include Midjourney, Firefly (Adobe), DALL-E (OpenAI), Llama (Meta), and Stable Diffusion (Stability AI). While many dominant models are proprietary, Stable Diffusion is an open-source model that was released to the public in August 2022 (Stability AI, 2022) and is used by a variety of generative AI tools, including Lensa (Prisma Labs, n.d.).

There has been considerable hype around visual generative AI tools, with claims that they are augmenting (Eapen et al., 2023), reimagining (Adobe Communications Team, 2023), and transforming (Roose, 2022) creativity and creative work. Indeed, the application of visual generative AI tools across a variety of fields including the visual arts, advertising, and film & TV has caused both celebration (e.g., Roose, 2022) and consternation (Equity, n.d.; Scherer, 2024). While some uses of visual generative AI tools have been decied as "inappropriate, unprofessional and disrespectful to audiences" (Sun and Harmon, 2024), others have been heralded as uses of "techno-vernacular creativity" that facilitates creative engagement within historically marginalized communities (Gaskins, 2022).

However, like many other AI technologies (e.g., Gillespie, 2024), visual generative AI models have been found to reproduce a series of harmful biases related to any number of identity characteristics, including gender (Sandoval-Martin and Martinez-Sanzo, 2024), race (Offert and Phan, 2022), mental health disorders (King, 2022), and religion (Alfano et. al, 2024). Quite often, these biases and stereotypes are imbricated with each other: Bianchi et al. (2023) found that Stable Diffusion generated "dangerous racial, ethnic, gendered, class, and intersectional stereotypes" that illustrated the "vast potential" of these models for "propagating harm along many axes of demographic identity" (p. 1494).

While it can be challenging to identify the root cause of these biases and stereotypes within a diffusion model (Luccioni et al., 2023), a model's training data has a significant impact on what it can do and how successfully it can do it (Buschek and Thorp, n.d., see also Denton et al., 2021). Midjourney, Stable Diffusion, and "perhaps hundreds" of other commercial models have been trained on LAION-5B (Buschek and Thorp, n.d.), an "uncurated" open-source dataset of over 5 billion image-text pairs that contains "strongly discomfoting and disturbing content for a human viewer" (Beaumont, 2022) including child sex abuse material (Thiel, 2023). An earlier version of this dataset (LAION-400) was found to have "troublesome and explicit images and text pairs of rape, pornography, malign stereotypes, racist and ethnic slurs, and other extremely problematic content" (Birhane, Prabhu, and Kahembwe, 2021). While LAION offers a disclaimer that LAION-5B should not be used in "ready to go industrial products", this warning has been largely ignored (Buschek and Thorp, n.d.) with clear consequences, as the Lensa case study illustrates.

3. Controversies & closures in AI media coverage

Scholarship on the influence of the news media has long established that media coverage about technology heavily influences how the public understands emerging technologies, as well as their risks and benefits; this is particularly true when audience members may not be familiar with a particular issue (Scheufele and Lewenstein, 2005). Chuan et al. (2019) argue that research evaluating how the media

discusses AI is essential to understand how public opinion about it is shaped; furthermore, as Brause et al. (2023) have noted, the media plays an important role in the dissemination of information about AI, shaping public perceptions and influencing public debates where “certain understandings of AI are advanced, and different stakeholders critically engage with the technology and its role in society, contributing to the communicative construction of the technology” (p.1). Ouchchy et al (2020) explain that it is particularly important to understand how the media portrays issues of AI and ethics, as they could have significant implications for AI development and regulation:

Because the members of the general public, as both consumers in the market economy and constituents of a liberal democracy, are key stakeholders for technology adoption—and, to a certain extent, for public policy and regulatory oversight—public opinion could affect what kind of AI is developed in the future and how AI is regulated by the government. (p. 927)

The representation of AI in the global media is a developing area of research, with approximately 30 empirical studies published since 2017 (Brause et al., 2023). These studies suggest that AI is usually discussed in the context of business or economics, and that the valence of this coverage is overwhelmingly positive (Brause et al., 2023; Chuan et al., 2019; Fast and Horvitz, 2017). In this sense, the media discourse surrounding Lensa is a bit of an anomaly, as it focuses more on social issues and includes a significant amount of critical coverage. However, as Fast and Horvitz (2017) and Chuan et al (2019) have pointed out, concerns about AI have also increased in news coverage in recent years, particularly in relation to AI ethics, the negative impact of AI on work, privacy, and loss of control over AI.

Dandurand et al (2023) note that the media has “become a powerful site of discourse formation in which certain voices and tropes about AI are authoritatively put forth while others are not” (p.2). Through this process, the media “participates in making, or not making more exactly, AI controversial” (p.3). Marres (2015) notes that in Science and Technology Studies (STS), controversy studies illustrates the cozy relationship between the formulation of knowledge claims and the organization of political interests (p. 656). The news coverage of Lensa certainly appears controversial in its discussion of data exploitation, misogyny, and racism in particular; these hot-button issues are often framed in a polemical way that seems to foment public discussion about the treatment of marginalized groups within algorithmic systems. However, what is notable about the Lensa coverage is that instead of providing novel pathways for public debate about these concerns and how to address them, generative AI harms are discussed in a way that is fossilized, in that it is frozen in place and stuck in the past (Dihal, 2024). Dandurand et al. (2023) ask, “what to do when AI’s coverage involves so much consensus about its benefits?”; conversely, the question that the Lensa coverage raises is: what to do when AI’s coverage involves so much consensus about its harms?

It is here where the communication studies concept of discursive closure can provide some assistance. Derived from Habermas’s concept of systematically distorted communication, Stanley Deetz (1992) theorizes discursive closure as existing “whenever potential conflict is suppressed” (p. 187). Discursive closure can take place through a variety of communicative processes and practices that “shut down or close off options for thinking otherwise” (Markham, 2021, p. 392). Such processes include disqualification (the exclusion of individuals from a discussion or conversation), naturalization (the reification of social dynamics and structures), neutralization (the positioning of value-laden activities as if they were neutral), and topical avoidance (Deetz, 1992; Leonardi and Jackson, 2003).

Markham (2021) suggests that discursive closure allows us to see how “certain patterns of thought, talk, actions, or interactions tend to function like negative feedback loops in social ecologies, discouraging evolution and change” (p. 392). Deetz (1992) argues that one way that discursive closure is achieved is “through the privileging of certain discourses and the marginalization of others” (p. 187), and that such discursive practices can either contribute to “further exploration of the subject matter” or “divert, distort, or block the open development of understanding” (p. 188). This curtailment of understanding and open discussion often has significant implications and consequences, usually in maintaining the status

quo: Deetz argues that when discussion of a particular issue is curtailed, “a particular view of reality is maintained at the expense of equally plausible ones, usually to someone’s advantage.” (p. 188)

Leonardi and Jackson (2003) have illustrated how discursive closure can deliver economic, political, and social advantages to certain ideologies over others by eliminating alternative conceptualizations of how things could have taken place or been done (p. 615). In their study of high-tech mergers in the late 1990s, they show how leaders of high-tech organizations invoked the “inevitability” of technology to justify managerial decisions to the public. Markham (2021) also found that discourses of inevitability were present in how members of the public engaged conceptually with digital platforms and algorithmic technologies, where alternatives to the status quo seemed “unimaginable” and any radical alternatives dismissed as improbable. She argues that discourses of inevitability teach us “that our present and possible futures are being determined by technology” and that even if the current trajectory suggests that “the future world is likely to be dystopian”, we are devoid of choice, because we are faced with an “either/or proposition” where we are either connected to technology or we do not exist (p. 397). As will be illustrated below, the theme of inevitability was both implicit and explicit in the discourse surrounding Lensa, foreclosing any significant debate or discussion of what AI should be and how it should operate.

4. Research design & methodological considerations

The study corpus consisted of 135 unique articles from the global English-language syndicated and tech press (see Figure 2). Articles were collected through two complementary methods. First, I searched Nexis for the term “Lensa AI” in English-language coverage from November 2022 to July 2023 across the system-defined regions of Africa, Asia, Europe, Latin America, Middle East, North America, and Oceania. This returned just under 400 results from 12 countries: Australia, Canada, Egypt, India, Mexico, Nigeria, Pakistan, Singapore, Switzerland, United Arab Emirates, United Kingdom, and United States. These results were refined by excluding content that was topically irrelevant. I then eliminated any duplicated or syndicated articles—of which there were many—to ensure that each article was only represented once in the corpus. This deduplication eliminated some countries from the analysis (i.e., Nigeria). After eliminating any irrelevant, duplicated, or syndicated articles, I was left with a total of 83 articles from Nexis. I then took a “snowball” approach (see Braithwaite, 2016) to the content linked or referred to in the Nexis results, adding any new articles that were specifically about Lensa. I then read these articles and followed any links in those pieces until I reached saturation and failed to come across any new material. I added a total of 52 articles through this process.

I analyzed the corpus using reflexive thematic analysis (Braun & Clarke, 2019; Byrne, 2020). I created a series of initial codes that reflected content that appeared repeatedly in the corpus; some examples of these codes include “skin tone”, “cleavage”, “celebrity”, “selfies”, “nudity”, “artist concerns”, “plastic surgery” and so on. These codes were then combined into broader themes, such as “sexualisation”, “privacy”, “narcissism”, and “training data”; these themes were then further condensed into the three overarching themes discussed below. I also reflected on the types/categories of coverage that were present in the sample: as will be discussed in the analysis, there were two types of coverage: coverage that focused on explaining what Lensa was (explainer coverage), and coverage that focused on critiquing Lensa (scandal coverage). After the topical themes and coverage categories were established, I returned to the corpus and reviewed each article, tallying which themes appeared within it and noting what coverage category it belonged to. This allowed me to note the prevalence of each theme and type of coverage within the corpus overall as well as any topical trends across countries/regions. It is worth noting that while articles were labelled as being predominantly one coverage type (explainer or scandal), some articles reflected multiple themes.

Table 1. Article totals

Total number of articles by country, as well as representative publications from each nation

	<i>Totals</i>	<i>Percentages</i>	<i>Sample Outlets</i>
<i>USA</i>	<i>46</i>	<i>34%</i>	<i>Techcrunch, Wired, MIT Technology Review, The New York Times</i>
<i>Mexico</i>	<i>42</i>	<i>31%</i>	<i>CE Noticias Financieras</i>
<i>UK</i>	<i>19</i>	<i>14%</i>	<i>The Independent, The Sun, The Guardian, Yorkshire Evening Post</i>
<i>India</i>	<i>17</i>	<i>13%</i>	<i>The Times of India, The Telegraph, Hindustan Times, Indian Express</i>
<i>Australia</i>	<i>3</i>	<i>2%</i>	<i>The Age, The Guardian</i>
<i>Switzerland</i>	<i>2</i>	<i>1%</i>	<i>Handelszeitung</i>
<i>Canada</i>	<i>2</i>	<i>1%</i>	<i>Toronto Star</i>
<i>UAE</i>	<i>1</i>	<i>1%</i>	<i>Wknd Magazine</i>
<i>Singapore</i>	<i>1</i>	<i>1%</i>	<i>The Straits Times</i>
<i>Egypt</i>	<i>1</i>	<i>1%</i>	<i>The Egypt Independent</i>
<i>Pakistan</i>	<i>1</i>	<i>1%</i>	<i>Pakistan and Gulf Economist</i>
<i>Total</i>	<i>135</i>	<i>100%</i>	

Nexis indicated that there were a few hundred results for “Lensa AI” in languages including Turkish, Spanish, French, German, Dutch, Arabic, and Portuguese. A cursory review of the French and Spanish language results suggests that some of the themes that were present in the English coverage are also present in other languages. At least part of the reason for this is because some English language news sources are translated into other languages and syndicated: for example, I found a Spanish-language article that turned out to be a translated version of a WIRED article that was already in the corpus. Of course, it is possible that there are themes present in other languages that aren’t reflected in this analysis, but it was notable that much of the coverage in this corpus was syndicated, particularly from US-based sources. This was also the case in other English-speaking countries: in Australia, for example, I only found three unique articles, two of which were from The Guardian; even when searching directly on key news sites like the Sydney Morning Herald to ensure I hadn’t missed anything significant, I only found syndicated materials from e.g., The Washington Post and the Wall Street Journal that were already included in the corpus. Furthermore, it is worth noting that there were more articles (and more countries) represented in the corpus before I traced syndicated articles to their sources and eliminated the duplicates. This suggests that the discourse about Lensa likely has a wider global audience than the countries reflected in this analysis, although the syndicated content clearly reflects the material included in this analysis. The issue of syndication and other factors related to the global news industry will be addressed more thoroughly in the Conclusion, but it is also a methodological consideration for this study.

Overall, my goal with this analysis was to offer an “illustrative rather than exhaustive account” (Braithwaite, 2016), of the discourse about Lensa, demonstrating its most salient features (p. 3). As the next section will illustrate, the discourse surrounding Lensa had distinct contours that appeared with

notable consistency across the countries represented in the corpus, albeit in different proportions. This clarity and consistency suggest that this analysis offers a clear illustration of the key issues associated with visual generative AI— and how they are discussed in the press— at this particular moment.

5. “Sex, art theft, and privacy”: Lensa in the global media

Although there were 11 countries represented in the corpus, there were four countries that comprised the majority of English-language Lensa coverage: the United States, Mexico, the United Kingdom, and India. The United States and Mexico dominated the coverage, with 34% and 31% of the corpus, respectively. News items from the United Kingdom comprised 14% of the sample, and Indian stories were 13%. Collectively, these four countries made up 92% of the corpus. The remaining 7 countries (Australia, Canada, Egypt, Pakistan, Singapore, Switzerland, and the UAE) collectively composed 8% of the corpus. While the dominance of the US, Mexico, the UK, and India within the sample likely has many underlying factors, one aspect could be related to the popularity of the app in each country: Lensa was the most downloaded app in the US in December 2022 (Silberling, 2022), and it was 6th on the charts in India (Rekhi, 2022).

The Lensa discourse had an intense burst of activity in the first two weeks of coverage, followed by a long tail: Lensa launched its Magic Avatar feature in late November 2022, the first article in the corpus was published in Mexico on November 29, 2022, and the last article in the corpus appeared in India on July 27, 2023.¹ However, 81% of the coverage in the corpus (110 articles) appeared within a month from when the first article was published; given that the corpus does not reflect any duplicate or syndicated content (which would increase those numbers considerably), that is a significant amount of coverage in a condensed period of time. As the chart below indicates, there were two spikes in coverage during this time: the first was on December 8, shortly after articles in TechCrunch and WIRED revealed that Lensa was generating explicit content, and the second spike was on December 13, the day after Melissa Heikkila (2022b) from MIT Technology Review published an article discussing the sexualized and racialized nature of her Lensa avatars.

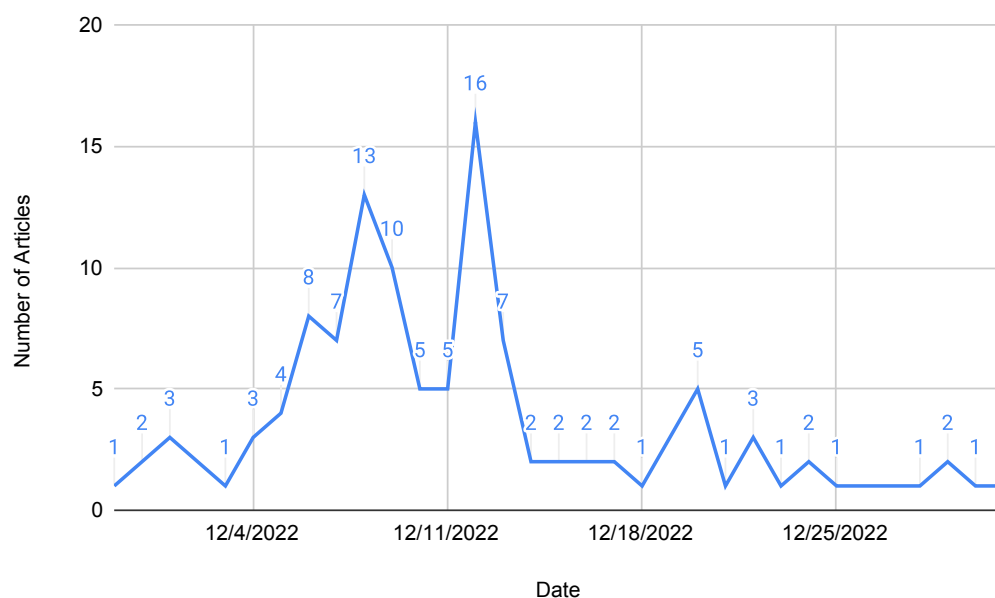


Figure 2. Article frequency by publication date

¹ It is possible that there are more articles that have appeared after July 2023, but that is the point at which data collection concluded

As mentioned in the Methodology section, there were two distinct types, or genres, of content in the Lensa discourse: scandal coverage and explainer coverage. Scandal coverage, which comprised 42% of the corpus, focused on perceived wrongdoing or norm violations related to the app and its use. Scandal coverage began to appear in the corpus on December 5. While scandal coverage was found in all countries aside from Egypt, Pakistan, UAE, and Singapore,² the topics that dominated the scandal-related articles came from a handful of articles (e.g., Hatmaker, 2022; Heikkila, 2022b; Kamps, 2022; Snow, 2022) published in US-based, agenda-setting tech publications like TechCrunch, MIT Technology Review, and WIRED. These articles were published within the first two weeks of the news cycle and were very influential; not only did they shape the discourse on a global level, but they were frequently linked to in other publications.

The other type of coverage was explainer coverage, which constituted 58% of the corpus and focused on describing what Lensa is and how it worked. Explainer coverage was present in the corpus from November 29th; indeed, all of the articles from the first six days of the news cycle fell into this category. Explainer coverage typically came from mainstream news publications. It contextualized Lensa as a viral phenomenon, talking about how “everyone” (or in some cases, specific celebrities) were using it. It sometimes included explanations of other generative AI apps, whether they were image-based (e.g., Remini, Midjourney, Firefly, DALL-E) or text-based (e.g., Chat GPT, Otter, Bard). This coverage typically explained the Magic Avatar feature, how to download and use the app, how much it cost, and if the results were any good. Some explainer coverage did this in a completely uncritical way, praising Lensa’s innovative nature and talking about how “cool” the generated images were. Most explainer coverage, however, tended to summarize or mention the key issues referenced within the scandal coverage as part of its description of the app as a hot new trend.

As noted previously, the discourse is remarkably consistent across contexts in terms of its content; in addition to the two categories of coverage, there were three overarching themes across the sample. The first theme focused on Predatory Data Practices. Articles related to this theme discussed unethical, exploitative practices relating to the non-consensual collection and use of training data, particularly from artists and app users. The second theme was Biased Content. Articles relating to this theme referred to the gender and racial biases reflected in some of the images generated by Lensa, including the sexualization of women, the lightening of skin tones, and the generation of explicit content. The third theme related to User Behavior. The content in this theme described how people used the app in specific ways, including celebrities and political figures. Below is a table that illustrates the distribution of each type of coverage and theme within the sample (Table 2).

² The lack of scandal coverage in these countries is likely attributable to the fact that there was only one article in the corpus for each of them.

Table 2. Article Distribution*Distribution of coverage type and themes across countries within the sample*

	USA	%	MEX	%	UK	%	IND	%	AUS	SWZ	CAN	EGY	PAK	UAE	SIN	Total	Total %
<i>Explainer</i>	19	41%	32	76%	7	37%	14	82%	1	2	0	1	1	1	1	79	59%
<i>Scandal</i>	27	59%	10	24%	12	63%	3	18%	2	0	2	0	0	0	0	56	41%
	46		42		19		17		3	2	2	1	1	1	1	135	
<i>Predatory Data Practices</i>	35	76%	15	36%	12	63%	9	53%	3	0	1	0	0	0	0	75	56%
<i>Biased Content</i>	30	65%	8	19%	12	63%	9	53%	1	2	1	0	0	0	0	63	47%
<i>User Behavior</i>	18	39%	18	43%	9	47%	4	24%	1	0	0	0	0	0	0	50	37%

5.1 Predatory data practices

The Predatory Data Practices theme was the most prevalent theme in the corpus, reflected in 56% of the articles and within six countries (USA, Mexico, UK, India, Australia, and Canada). It was most commonly discussed in the US and UK articles (76% and 63%, respectively), with a little more than half of the Indian articles (53%) and one-third of the Mexican articles (36%) referencing data-related issues.

This theme focused predominantly on two distinct, but related topics: the use of art as training data in the Stable Diffusion model and Lensa’s privacy policy, which stated at the time that users’ photos could be used to train the app. Coverage reflecting this theme largely focused on the unethical or non-consensual use of data in AI technologies and the implications of those practices for both specific communities and users in general. While these practices were heavily critiqued for the distress and harm that they caused, they were also largely framed as unavoidable. This aligns with Markham’s (2021) observation that there is an “overall expectation” that technology companies collect and sell (personal) data and that an alternative to this state of affairs seems unimaginable (pp. 390-391).

5.1.1 “Robot replacements”: AI and/as art

The discussion of art and artists within the Lensa discourse originated in the US, but was also the focus of articles in Canada, Australia, and the UK. It was not the first time a debate about art, the ethics of training data, and the devaluation of creative labor had appeared in the media; rather, the launch of Lensa “reignited discussion” (Sung, 2022) of these issues.

Within the corpus, there was a common framing that Lensa was “stealing” from artists. This phrase first appeared in the corpus in a TechCrunch (US) article by Taylor Hatmaker (2022) on December 5, noting that “for every 10 Lensa avatars there’s one Cassandra in the comments scolding everyone for paying for an app that steals from artists”. In an NBC News (US) article on December 7, (Sung, 2022), artist Karla Ortiz commented, “Companies like Lensa say they’re “bringing art to the masses, but really what they’re bringing is forgery, art theft [and] copying to the masses.” Within a week of Hatmaker’s original article, this topic appeared in articles across the globe.

The “theft” that Lensa was seen to perpetuate— described in one article as “arguably the biggest art heist in history” (Tran, 2022)— had three intertwined elements: the unethical use of art as training data, a lack of fair compensation to artists, and the replacement of artists by AI. This was discussed as a multi-layered insult: artists’ work was used without consent— and without remuneration— to train models which were duplicating their styles³ and stealing jobs from them.⁴ These concerns were summarized by artist Molly Crabapple (2022) an op-ed in the Los Angeles Times (US), who stated,

These data sets were not ethically obtained. LAION sucked up 5.8 billion images from around the internet, from art sites such as DeviantArt, and even from private medical records...They took it all without the creator’s knowledge, compensation, or consent. Once LAION had scraped up all this work, it handed it over to for-profit companies — such as Stability AI, the creator of the Stable Diffusion model — which then trained their AIs on artists’ pirated work... AIs can spit out work in the style of any artist they were trained on — eliminating the need for anyone to hire that artist again.

Similar claims were repeated throughout the corpus. In *The Independent* (UK), Josh Marcus (2022) noted that the AI models used by apps like Lensa “harvest the stylistic DNA of individual artists, then allow strangers to borrow elements from their work without offering any credit” in an “incredibly direct” manner. This was echoed by Australian artist Kim Leutwyler, who explained to *The Guardian* (AUS) that “some artists are having their exact style replicated exactly in brush strokes, colour, composition – techniques that take years and years to refine” (Kelly, 2022). While some articles also noted that AI models are seen by some artists as “a help, not a threat” (Marcus, 2022) and that something that “augments” the capabilities of artists (Flux, 2022), the majority of articles framed Lensa and its ilk primarily as “robot replacements” that “vampirized the work of artists” (Crabapple, 2022). Furthermore, this state of affairs was largely framed as a *fait accompli*, with artists being described as “exhausted and powerless” (Tran, 2022) with no recourse, because it is “unclear” how artists can protect their intellectual property “from being sucked into AI models” even if they try (Marcus, 2022). This sense of inescapable data collection was also present in the coverage that discussed Lensa’s privacy policy.

5.1.2 “You did not consent to this”: Lensa’s privacy policy

Concerns about Lensa’s privacy policy appeared around the same time as concerns about Lensa as an art “thief”: on December 5, articles in *Art News* (US) and *Refinery29* (UK) noted that Lensa’s terms and conditions allow it to “use the manipulated photos you upload for any way it sees fit” (Roberts, 2022). In an article titled “Careful — Lensa Is Using Your Photos to Train Their AI”, Shanti Escalante-DeMattei (2022) noted that “Lensa’s privacy policy and terms of use stipulate that the images users submit to generate their selfies, or rather the ‘Face Data,’ can be used by Prisma AI, the company behind Lensa, to further train the AI’s neural network.” The privacy policy was brought up again on December 7 at the end of an explainer article in *The New York Times* (Kircher and Holtermann, 2022), where it was explained that users’ “face data” was deleted within 24 hours, but that any uploaded photos or videos could be used to train Lensa’s “algorithm”. The issue of Lensa’s privacy policy appeared regularly after that, particularly in explainer coverage, where it was framed as an issue of “safety”, “risk” or “danger”, especially in India, Mexico, and the UK.

Similar to the art-related coverage, articles discussing Lensa’s privacy policy often made connections to the broader implications of user data being used to train AI systems. In the *New York Times* explainer article referenced above, Jen King, a privacy and data policy fellow at Stanford, suggested that it was user data that was Lensa’s profit motive, stating, “I doubt that the whole business model is, ‘Give us \$10 or \$15 and we’ll send you back an A.I. glam shot’” (Kircher and Holtermann, 2022). In a December 10 *Washington Post* article, Shira Ovide (2022) argued that Lensa and other “cool AI technologies” are “built on all the information we’ve put out into the world”, which is having significant unintended consequences: “that drunk selfie you posted on Instagram” is now “training fuel for an artificial

³ Fantasy landscape artist Greg Rutkowski has become a poster child for this issue, as his name became one of the most used prompts when Stable Diffusion was released in August 2022 (See Heikkilä 2022a).

⁴ For an example of this, see Maimann, 2022.

intelligence system that helps put an innocent person in jail.” While these outcomes are framed as unfortunate, they are also framed as unavoidable: as Ovide put it, “once you put digital bits of yourself or your loved ones online, you lose control of what happens next.”

Discussions of Lensa’s privacy policy also often harkened back to previous privacy scandals associated with apps like Faceapp (mentioned in US and Indian coverage), Snapchat filters, and Reface (both mentioned in UK coverage). This is not entirely surprising, given that before its Magic Avatar feature was launched, Lensa was exclusively an image-editing app. Image-editing apps were also invoked in discussions relating to specific types of aesthetics and beauty standards, such as lightened skin tones and normative body sizes. These kinds of biases were a significant topic of discussion on their own, particularly in the scandal coverage.

5.2 Biased Content

Biased Content was the second most prevalent theme in the corpus, reflected in 47% of articles. However, it was arguably the most impactful: the two spikes in coverage on December 8 and December 13 were related to articles that focused on this theme and had a big influence on its frequency in the coverage. The presence of the Biased Content theme was relatively even across US (65%), UK (63%) and Indian (53%) coverage. However, only 19% of the Mexican coverage referenced it, which featured predominantly in scandal coverage. The Biased Content theme also appeared in coverage from Australia and Canada, and the two articles from Switzerland focused exclusively on it.

The content in this theme largely focused on two topics: Lensa’s generation of explicit images that sexualized women and girls, and the fact that it created images that were racialized in specific ways. These topics were often intertwined, with some of the most influential articles engaging with both topics in a way that reflects the intersectionality of these concerns. The discussion of how women— and women from minoritised racial groups in particular— are represented by Lensa reflects a broader discourse about the biases of algorithmic technologies (e.g., Noble, 2018). However, unlike this discourse, which understands algorithmic bias as a multidimensional problem, the Lensa coverage laid the blame for Lensa’s production of explicit and biased imagery almost exclusively at the feet of the Stable Diffusion model and the “unfiltered internet content” (Prisma Labs, n.d.) that it was trained on. While the sexualized and racialized images Lensa produced were heavily critiqued by the press, they were also portrayed as an unfortunate— but perhaps unavoidable— reflection of the “biases humans incorporate into the images they produce” (Prisma Labs, n.d.).

5.2.1 “Why do all my AI avatars have huge boobs?”⁵: Explicit and sexualized content

The topic of explicit and/or sexualized content first appeared in the corpus on December 5 in a Refinery 29 (UK) article that noted that Twitter users had been complaining about the fact that Lensa had “slimmed down larger-bodied people” and perpetuated “the male gaze” through “hyperfeminised ideations of female-identifying individuals” (Roberts, 2022). The complaints about Lensa sexualizing women continued on December 6 in an op-ed in Business Insider (Hill, 2022) that explained how a Lensa user with a history of disordered eating was triggered by the “heroin chic” images that the app generated, making her appear “model-thin, sexualized, and miserable.”

The issue of sexualization was further amplified by articles in TechCrunch (US), WIRED (US), and The Guardian (UK) that discussed how Lensa produced explicit content. The TechCrunch article, titled “It’s way too easy to trick Lensa AI into making NSFW images”⁶ also came out on December 6. Author Haje Jan Kamps (2022) noted that, even though Stable Diffusion has an explicit content filter in place, images that have been modified by Photoshop somehow deactivate that filter and “gladly churn out a

⁵ Mercado, M. (2022, December 12). Why Do All My AI Avatars Have Huge Boobs? *The Cut*. <https://www.thecut.com/2022/12/ai-avatars-lensa-beauty-boobs.html>

⁶ NSFW is an acronym for “not safe for work”, aka explicit or pornographic content

number of problematic images”. He then highlighted the implications of being able to “create near-photorealistic AI-generated art images by the hundreds without any tools other than a smartphone, an app and a few dollars”, particularly in relation to the generation of AI-generated, or “deepfake”⁷ pornography, both for celebrities and normal people alike.

The following day, WIRED published an article titled “‘Magic Avatar’ App Lensa Generated Nudes From My Childhood Photos” (Snow, 2022). Researcher Olivia Snow (2022) detailed how, when she tested Lensa’s “no kids, adults only” policy by submitting a combination of “mousy” adolescent and “awkward” childhood photos, Lensa produced “fully nude photos of an adolescent and sometimes childlike face but a distinctly adult body.” Snow explained that she had been inspired to do this because many users (but “primarily women”) had pointed out that the app “ascribes cartoonishly sexualized features, like sultry poses and gigantic breasts, to their images” and generated “fully nude results despite uploading only headshots”, which in turn made people feel “very violated”. The Guardian (UK) ran a similar experiment, feeding Lensa images of “feminist icons” Amelia Earhart, Betty Friedan, and Shirley Chisholm, and receiving results comparable to Snow’s:

The author of *The Feminine Mystique* became a nymph-like, full-chested young woman clad in piles of curls and a slip dress. Chisholm, the first Black woman elected to US Congress, had a wasp waist. And the aviation pioneer was rendered naked, leaning on to what appeared to be a bed (Demopolous, 2022).

Lensa’s production of erotic imagery was primarily attributed to the LAION training dataset used by Stable Diffusion and other AI models. Articles in the Wall Street Journal (O’Brien, 2022) and MIT Technology Review (Heikkilä, 2022b) explained that because the internet is rife with images of naked or scantily clad women, the dataset that Stable Diffusion was trained on is skewed toward sexualized images, whether women want to be depicted in that manner or not. A spokesperson from Lensa’s parent company explained to the MIT Technology Review (Heikkilä, 2022b) that “the man-made, unfiltered online data introduced the model to the existing biases of humankind” and as such, they couldn’t possibly “consciously apply any representation biases or intentionally integrate conventional beauty elements” into the app and its output. In other words, the ultimate blame for Lensa’s production of sexualised imagery lies with humans and their biases, and not with the technology itself. This was an explanation that was also provided to explain Lensa’s production of images that reflected racial biases in addition to gender biases.

5.2.2 “Generic hot white girl”: Racialized images

The issue of racialization first appeared in the corpus in the same Art News article that highlighted the concerns with Lensa’s privacy policy (Escalante-DeMattei, 2022). It noted that writer Maya Kotomori had received a batch of avatars that made her look white, despite the fact that she is a “fair-skinned black person”. AI For The People founder Mutale Nkonde explained in The Washington Post (US) that the lack of representation of dark-skinned people in AI training images means that AI technologies like Lensa have trouble analyzing and reproducing images of them; as a result, Black women get images that look like a “generic hot white girl” (Hunter, 2023). This phenomenon was also described by Swiss writer Rebecca Stevens Alder (2022), who suggested that the “white privilege and racism” built into Lensa meant that it “did not know what to do when faced with a Black face”. Polygon’s Nicole Clark (2022), who is Taiwanese and white, similarly said that Lensa had “no idea what to do with my face” and created images that looked alternately East Asian or white, but like “complete strangers, save for one particular quirk or other of mine, like my jawline or eye shape”. Clark noted that whether she looked white or East Asian depended on the filter she chose, with the images for “Kawaii” looking more Asian and the images for “Light” and “Fantasy” looking white. She explained that while one image in the “Cosmic” set gave

⁷ Deepfakes are “hyper-realistic videos that apply artificial intelligence (AI) to depict someone say and do things that never happened” (Westerlund, 2019, p. 38); AI-generated images can also be considered deepfakes.

her “random cleavage”, she did not receive any nude photos; however, that was certainly not the case for other people with Asian ancestry.

The imbrication of sexualization and racialization in the results for people from minoritized racial groups was discussed at length in a viral article by Melissa Heikkila (2022b) of the MIT Technology Review that was published on December 12. Heikkila, who described herself as mixed race, stated that her Lensa avatars were “cartoonishly pornified” and noted that out of a total of 100 avatars, 14 were topless and a further 16 were in “skimpy” outfits and sexualized poses. She suggested that this extreme sexualization was attributable to her Asian heritage, which seemed to be the dominant feature that the AI model recognized.

I got images of generic Asian women clearly modeled on anime or video-game characters. Or most likely porn, considering the sizable chunk of my avatars that were nude or showed a lot of skin. A couple of my avatars appeared to be crying. My white female colleague got significantly fewer sexualized images, with only a couple of nudes and hints of cleavage. Another colleague with Chinese heritage got results similar to mine: reams and reams of pornified avatars. Lensa’s fetish for Asian women is so strong that I got female nudes and sexualized poses even when I directed the app to generate avatars of me as a male.

Snow (2022) also noted in her article that “a woman of Asian descent” explained to her that in photos where she didn’t look white, she was given “ahegao face”, a facial expression used in erotic manga to “highlight a hyperintense orgasm” (Santos, 2020).

Like the sexualization issue, the blame for Lensa’s racialization problems within this coverage is laid at the feet of existing structural inequalities: society is biased, consequently, so are the images that they create; if these are the images that AI models are trained upon, such content is inescapable and simply the cost of engagement. For users who receive inappropriate or biased content, what are the options? According to Glamour (UK)’s Anya Meyerowitz (2022), there is only one: don’t take part. “It’s clear that the app is violating our rights, perpetuating misogyny and is based on, at best, shaky ethics”, Meyerowitz concludes. “We’d encourage everyone, not just women, to stop using the app.”

5.3 User behavior

The idea that users should stop using Lensa is a common subtext in the User Behavior theme, although unlike the Predatory Data Practices and Biased Content themes, it is our own narcissistic tendencies– and not necessarily the AI models– that we should be eschewing.

The User Behavior theme encompassed content that discussed how specific people or groups were using Lensa. This included both average users and celebrities in both explainer and scandal content. User Behavior was the least prevalent theme in the American (39%), British (47%) and Indian (24%) coverage. It was, however, the most prevalent theme in Mexican coverage (43%), largely due to the volume of articles that mentioned how specific public figures shared their Magic Avatars on Instagram, including politicians (e.g., Mexican President Claudia Sheinbaum, Senator Ricardo Monreal) and celebrities (e.g., commentator and host Marisol Gonzalez, singer Yahaira Plasencia). Most of the User Behavior content in Mexico was explainer coverage, but there was also some Scandal content that talked about the influence of artificial intelligence on our self-perception. In an article titled “Looking like a cyborg to look pretty: how artificial intelligence is shaping the canon of beauty” (Bou, 2022), the author argues that the popularization of Lensa and other image-focused technologies promotes an unrealistic, homogenous, “robotic beauty” that reflects the subordination of our identities to the dictates of technological advances.

The relationship between identity and visual technologies like Lensa was a dominant thread in the User Behavior theme. An article in Time (US) suggested that there was a “vanity factor” in play to Lensa’s appeal, and a New York Times article titled “AI’s Best Trick Yet Is Showering Us With Attention” (Haigney, 2023) proposed that apps like Lensa are popular because they engage our narcissistic tendencies. “The fundamental appeal of apps like this is, of course, to our own self-involvement,” author Sophie Haigney argued, concluding that “the app tricks me into feeling seen, but really it is just me, trying once again to see myself.” An article in The Guardian (UK) (Demopolous, 2022) made a similar

argument, suggesting that the “allure” of Lensa was based in a self-oriented “curiosity”. Even though Lensa often produced images of “teenage-boy comic-book fantasy girls” that made her feel “icky”, illustrator Cat Willet explained that “even if you know it’s wrong, you still want to see what yours looks like”. While most articles framed this self-interest as a shallow, albeit relatively harmless, narcissism, other articles suggested that the “hot AI selfies” produced by Lensa might exacerbate body image concerns or disorders (Klee, 2022). Other articles suggested that engagement with Lensa might have even more extreme effects: an article from TMZ that was referenced in several articles in the corpus (“Celeb Surgeons Say”, 2022) reported that celebrity plastic surgeons were receiving visits from people requesting to look like their Lensa portraits.

There were some articles, however, that described Lensa being put to an empowering use: enabling transgender and non-binary people to experience “gender euphoria” through the production of images that align with their gender identity. In Refinery29 (UK), Millie Roberts (2022) discussed the “transformative power of avatars” for LGBTQ+ technology users to enhance “the lives of diverse and minority communities” by offering venues for identity exploration in online contexts. Adam Smith (2022) explained in Openly News (UK) that for users with gender dysphoria, their Lensa portraits could be “affirming” and Saskia Maxwell Keller (2022) of Out (US) described the “trans joy” inspired by Lensa as a “positive force”. However, Roberts, Smith, and Keller all noted that not all trans users had equally positive experiences, and that the controversies around Lensa’s training data made some trans users reluctant to engage with it.

Despite the relative novelty of Lensa’s Magic Avatar feature, the discourse that explains, analyzes, and critiques it does not offer a new perspective; instead, it treads well-worn paths that portray the negative aspects and consequences of visual generative AI as unavoidable or inevitable. The key themes of “sex, art theft and privacy” (Biron, 2023)—and, importantly, how they are discussed—in the Lensa discourse connect to existing framings and understandings of these issues. This has significant implications for how we understand the key problems inherent in AI technologies, as well as the solutions that we propose and pursue— or don’t— to address those concerns.

6. “The horse has left the barn”: Discursive closure and visual generative AI

Brause et al. (2023) have pointed out that media narratives and framings have a direct impact on policy making, economic investments, and research activities in emerging technological fields (p. 11), and Bareis and Katzenbach (2021) similarly argue that there is “a strong role for discourse in shaping the present and future sociotechnical pathways” of AI (p. 858). For these reasons, how the challenges and harms of visual generative AI are presented within the global news media have potentially immense consequences for how— or even if— these harms are remediated (see Shane, 2023).

The Lensa coverage suggests that, with few exceptions, there is a discourse of inevitability (Leonardi and Jackson, 2003; Markham, 2021) in play that fails to offer any novel or alternative framings— or even opportunities for discussion—to the harms posed by visual generative AI. The discourse about Magic Avatars certainly identifies those harms, and even strongly critiques them; however, it also treats them as intractable problems that are already too big to solve.

Although Lensa was arguably a novel technology at the time, it is remarkable how closely the press coverage about it hewed to existing framings and discourses about social media platforms and other types of algorithmic media. The discussions of predatory data practices closely align with popular discourses about users as exploited data subjects that took off exponentially in the wake of the Cambridge Analytical scandal (e.g., Isaac and Hanna, 2018; Koidl and Kapanova, 2020), the critiques of Lensa’s sexualized and racialized results map directly onto discourses about “algorithms of oppression” and other forms of gendered and racialized algorithmic biases (e.g., Noble, 2017; Benjamin, 2019), and analyses of Lensa user behaviors were plucked directly from the selfie moral panic playbook, with their insistence that

digital technologies amplify our most deleterious narcissistic tendencies, especially if “we” are women (Miltner and Baym, 2015; Senft and Baym, 2015; Abidin, 2016; Tiidenberg, 2018).

To be clear, algorithmic bias and predatory data practices are significant social concerns that are deserving of public attention (the gendered media panic around selfies and other self-representation practices, perhaps less so). However, it is not *that* these matters are being discussed but *how* they are being discussed that is the problem. By repeating existing framings of these issues, it reinscribes hegemonic understandings of them and perpetuates a sense of inevitability about the status quo. In other words, you might be upset that these technologies are stealing your data and making porn out of your photos, but there’s not much you can do about it. As Markham (2021) explains,

Focusing on how discourses are normalized or locked into repetitive loops helps specify how hegemony works in everyday practices. In systems of highly effective oppression or, what Gramsci labeled “control through consent,” people shut down alternatives themselves, naturalizing problems as “just the way things are.” (p. 392)

This framing of AI harms as “just the way things are” is present throughout the corpus, and particularly in the Predatory Data Practices and Biased Content coverage. This framing manifests in two key ways. The first is that AI simply reflects pre-existing societal biases: Prisma Labs spokesperson Anna Green explained to The New York Times that Lensa was not consciously applying biases, but rather that “essentially, A.I. is holding a mirror to our society” (Chen, 2023). This suggests that society— and not a series of corporate policy and design choices— is primarily responsible for these problems, an explanation that conveniently eschews the fact that generative AI systems have been proven to amplify and worsen societal biases (e.g., Nicoletti and Bass, 2023). The other framing tends to suggest that any potential interventions to address these concerns are simply too late. In describing the recourse available to people whose images are included in the LAION 5B dataset, TechCrunch’s Taylor Hatmaker (2022) explains that EU citizens can file takedown requests for individual images but “that’s about it”, because “the horse has already left the barn”. Shira Ovide (2022) of The Washington Post had a similarly fatalistic attitude about the creation of AI systems, stating that readers’ feelings about their data being used to train AI models were largely inconsequential, because they were powerless to stop it:

Good or bad, these AI systems are being built with pieces of you. What are the rules of the road now that you're breathing life into AI and can't imagine the outcomes? [...] Being part of the collective building of all these AI systems might feel unfair to you, or amazing. But it is happening.

Markham (2021) explains that inevitability and powerlessness can be outcomes of naturalization, which takes place when current elements of sociotechnical contexts are accepted simply as the way things are (p. 397). Even though Stable Diffusion and its ilk have been available to the public for less than two years, the idea that these models— and the systems they undergird— will inevitably cause harm seems to be already taken for granted. This is a phenomenon that Daniel Chandler (1995) refers to as the “technological imperative” (see also Widder et al., 2022). Chandler writes that “the technological imperative is a common assumption amongst commentators on 'new technologies'. They tell us, for instance, that the 'information technology revolution' is inevitably on its way and our task as users is to learn to cope with it.”

Indeed, the solution in the corpus suggested to those most harmed by visual generative AI was to simply abstain from using it if they didn’t like the images it was producing. However, given that AI is already incorporated into many key digital technologies, engaging in media refusal (Portwood-Stacer, 2013) is not always desirable or feasible, especially as it could result in exclusion from the public sphere for groups who are already excluded and marginalized, denying them important sources of community and resistance (e.g., Sobande et al., 2019). Furthermore, these kinds of logics place the onus on the user who is harmed, instead of on the technology (and corporations) that are responsible for the harm(s): if the solution for AI harms is for marginalized users is to refrain from using these systems, it offers a reprieve to those who own and control these technologies from having to address any underlying issues that they perpetuate.

In this way, we can see discursive closure at work. By acquiescing to a discourse of inevitability, the Lensa coverage offers “a particular view of reality that is maintained at the expense of equally plausible ones” (Deetz, 1992, p. 188). Consequently, any alternative solutions to the issues at hand end up being downplayed or ignored. In her Los Angeles Times op-ed, Molly Crabapple (2022) argued that “data sets such as LAION-5B must be deleted and rebuilt to consist only of voluntarily submitted work. AIs trained on copyrighted art must also be pulled.” This suggestion aligns with some recent work in computer science that recommends that “learning methods that physically manifest stereotypes or other harmful outcomes be paused, reworked, or even wound down when appropriate, until outcomes can be proven safe, effective, and just” (Hundt et al., 2022, p. 743). Some technology companies are already taking this approach: in February 2024, Google took Gemini, its newly-launched AI image generator, offline when it began producing offensive images (Griffin, 2024). And yet, these ostensibly reasonable ideas were found nowhere else in the press coverage of Lensa.

Markham (2021) explains that the power of anticipatory logics that flow through everyday discourse around technologies like AI end up building and reinforcing “a hegemonic ideology of external power and control” (p. 384). As a result, more expansive and/or creative approaches to the challenges and potential harms of emerging technologies end up being ignored or dismissed as unlikely, constraining the boundaries of possibility. Dandurand et al. (2023) suggest that the “indeterminacies and uncertainties” of novel technologies like visual generative AI can “become settled when a few experts close debates, obfuscate contrasting expectations, and shape the political economy of science and technologies” (p. 3). As the discourse about Lensa illustrates, this process is already happening in relation to visual generative AI.

7. Conclusion

This paper analyzed the global press coverage of visual generative AI app Lensa and found that the discourse surrounding it focused on the app’s predatory data practices, the biased content it produced, and the user behaviors associated with it. I argue that this coverage provides evidence of discursive closure around key issues associated with visual generative AI that supports the maintenance of the status quo. I suggest that the press coverage of Lensa, which both articulates key AI-related harms and frames those harms as intractable and insolvable, creates a discourse of inevitability that has implications for how these issues are understood by the public, and for the approaches that are taken to address them. In doing so, a more imaginative public discussion of what visual generative AI could– or should– be is foreclosed.

Markham (2021) argues that “people seem to have difficulty imagining futures in ways that do not reproduce current ideological trends or cede control and power to external, mostly corporate, stakeholders” (Markham, 2021, p. 385). Indeed, it can be difficult to imagine things differently if you are consistently provided with a similar set of limited visions over and over again; as Markham also points out, “this continuously repeated mantra of powerlessness to avoid what is inevitable is not just something we invent; it is learned (or taught) in micro doses through our news feeds” (p. 397). Part of the issue is how journalism– both a public service and a struggling industry– operates in the 21st century. As Ananny and Finn (2020) point out,

Today’s news happens not just where journalists are ready to see news happening, but where news infrastructures have been designed to look. Part of appreciating the myriad forces and values of the contemporary networked press means understanding how infrastructures of humans (journalists, designers, audiences) and nonhumans (data sets, algorithms, interfaces), together, construct ways of seeing social worlds and translating those visions into familiar forms of news (p. 1613)

The evolution of news into a “borderless global market” has shifted news production practices and media consumption habits, where major media outlets in Western countries are able to deliver content to international audiences at a low cost (Rafeeq and Jiang, 2018). Furthermore, the datafication of the

audience and the metrification of news coverage has transformed editorial and journalistic practices, with journalists choosing to write articles that they might not see as newsworthy but feel will perform well with audiences (Dodds et al., 2023). This is not to say that the issues raised in the Lensa coverage are not newsworthy; however, it might not be a huge leap to suggest that part of the reason that Lensa received so much attention in the press is because it was a topic that performed well, which contributed to widespread syndication and further coverage. It also might not be a huge leap to suggest that how Lensa was covered was likely influenced by previous framings or angles that had performed well in the past or would invariably do so; tech coverage is a beat for specific journalists and outlets, and it doesn't take a veteran editor to know that a story about a popular app making non-consensual pornography out of people's photos is certainly clickworthy.

However, while these journalistic circumstances and choices may be at least partial explanations as to why the Lensa discourse looks the way it does, it doesn't negate the problem posed by these determinist—and unimaginative—framings. As McKelvey et al. (2023) argued in the CFP for this Special Issue, AI scandals—like Lensa's generation of racialized, sexualized images combined with their exploitative data practices—may offer “easy, high-engagement stories” that are appealing to news outlets and audiences alike, but such news coverage also fails to offer meaningful public engagement, not to mention democratic praxis. Indeed, Deetz (1992) argues that “it should not be surprising that systems of domination are protected from careful exploration”. In the case of Lensa, it may seem counterintuitive that the extensive discussion of the downsides of visual generative AI would provide any distinct advantages to the corporate interests who dominate in this space, but in identifying these harms and then framing them as inexorable, the “careful exploration” of potentially inconvenient—and expensive—solutions is discarded. Identifying the problems of visual generative AI and then insisting they're unfixable is probably a more effective way to curtail meaningful discourse about these concerns than focusing exclusively on AI's benefits.

In the face of this, it may seem that altering the trajectory of the discourse about visual generative AI—not to mention its development and deployment—is unlikely or insurmountable. However, we need not yet wave the white flag; it is early days yet. The history of technology shows us that technological development—and how it is discussed—are by no means inevitable, and that things could always have been otherwise (e.g., Marx, 2010; Pinch and Bijker, 1984). While the challenges may be significant, change is not impossible with some imagination and a collective refusal to accept the status quo.

References

- Abidin, C. (2016). “Aren't These Just Young, Rich Women Doing Vain Things Online?": Influencer Selfies as Subversive Frivolity. *Social Media + Society*, 2(2), 2056305116641342. <https://doi.org/10.1177/2056305116641342>
- Adobe Communications Team. (2023, October 18). *Generation AI: Reimagining creativity with generative AI*. Adobe Blog. <https://blog.adobe.com/en/publish/2023/10/18/generation-ai-reimagining-creativity-with-generative-ai>
- Alder, R. S. (2022, December 17). The White Privilege And Racism Of Lensa's Magic Avatars. *ILLUMINATION-Curated*. <https://medium.com/illumination-curated/the-white-privilege-and-racism-of-lensas-magic-avatars-355b1b53d306>
- Alfano, M., Abedin, E., Reimann, R., Ferreira, M., & Cheong, M. (2024). Now you see me, now you don't: An exploration of religious exnomination in DALL-E. *Ethics and Information Technology*, 26(2), 27. <https://doi.org/10.1007/s10676-024-09760-y>
- Ananny, M., & Finn, M. (2020). Anticipatory news infrastructures: Seeing journalism's expectations of future publics in its sociotechnical systems. *New Media & Society*, 22(9), 1600–1618. <https://doi.org/10.1177/1461444820914873>
- Bareis, J., & Katzenbach, C. (2022). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values*, 47(5), 855–881. <https://doi.org/10.1177/01622439211030007>
- Beaumont, R. (2022, March 31). *LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS*. LAION. <https://laion.ai/blog/laion-5b>
- Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. John Wiley & Sons.
- Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *Proceedings of*

- the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1493–1504.
<https://doi.org/10.1145/3593013.3594095>
- Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). *Multimodal datasets: Misogyny, pornography, and malignant stereotypes* (arXiv:2110.01963). arXiv. <https://doi.org/10.48550/arXiv.2110.01963>
- Biron, B. (2023, January 16). *Sex, art theft, and privacy: Lensa exploded overnight — and now the avatar app is dealing with public backlash. Here's what you should know.* Business Insider. <https://www.businessinsider.com/lensa-ai-raises-serious-concerns-sexualization-art-theft-data-2023-1>
- Bou, C. P. (2023, February 12). Assemblar-se a una ciborg per estar guapa: Com la intel·ligència artificial està modelant el cànon de bellesa. *El Periodico*. <https://www.elperiodico.cat/ca/societat/20230212/bellesa-intelligencia-artificial-cara-ciborg-filtres-canons-82583384>
- Braithwaite, A. (2016). It's About Ethics in Games Journalism? Gamergaters and Geek Masculinity. *Social Media + Society*, 2(4), 2056305116672484. <https://doi.org/10.1177/2056305116672484>
- Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597. <https://doi.org/10.1080/2159676X.2019.1628806>
- Brause, S. R., Zeng, J., Schäfer, M. S., & Katzenbach, C. (2023). Media representations of artificial intelligence: Surveying the field. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 277–288). Edward Elgar Publishing. <https://doi.org/10.4337/9781803928562.00030>
- Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., & Chuan, C.-H. (2019). Artificial Intelligence and Journalism. *Journalism & Mass Communication Quarterly*, 96(3), 673–695. <https://doi.org/10.1177/1077699019859901>
- Brown, D. (2023, July 13). The Viral AI App That's Triggering Baby Fever. *Wall Street Journal*. <https://www.wsj.com/articles/want-to-see-your-future-babies-this-viral-app-can-show-you-f4f65bdd>
- Buschek, C., & Thorp, J. (n.d.). *Models All The Way Down*. Knowing Machines. Retrieved 2 October 2024, from <https://knowingmachines.org/models-all-the-way>
- Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56(3), 1391–1412. <https://doi.org/10.1007/s11135-021-01182-y>
- Celeb Surgeons Say Clients Asking to Look Like Lensa App AI Portraits*. (2022, December 12). TMZ. <https://www.tMZ.com/2022/12/12/plastic-surgeons-say-clients-asking-look-like-lensa-app-portraits/>
- Chandler, D. (1995). *Technological Determinism: The Technological Imperative*. Retrieved 20 May 2024, from <http://visual-memory.co.uk/daniel/Documents/tecdet/tcet07.html>
- Chen, B. X. (2022, December 21). How to Use ChatGPT and Still Be a Good Person. *The New York Times*. <https://www.nytimes.com/2022/12/21/technology/personaltech/how-to-use-chatgpt-ethically.html>
- Chuan, C.-H., Tsai, W.-H. S., & Cho, S. Y. (2019). Framing Artificial Intelligence in American Newspapers. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 339–344. <https://doi.org/10.1145/3306618.3314285>
- Clark, N. (2022, December 20). Lensa's viral AI art creations were bound to hypersexualize users. *Polygon*. <https://www.polygon.com/23513386/ai-art-lensa-magic-avatars-artificial-intelligence-explained-stable-diffusion>
- Crabapple, M. (2022, December 21). Op-Ed: Beware a world where artists are replaced by robots. It's starting now. *Los Angeles Times*. <https://www.latimes.com/opinion/story/2022-12-21/artificial-intelligence-artists-stability-ai-digital-images>
- Dandurand, G., McKelvey, F., & Roberge, J. (2023). Freezing out: Legacy media's shaping of AI as a cold controversy. *Big Data & Society*, 10(2), 20539517231219242. <https://doi.org/10.1177/20539517231219242>
- Deetz, S. (1992). *Democracy in an age of corporate colonization: Developments in communication and the politics of everyday life*. SUNY press.
- Demopoulos, A. (2022, December 9). The inherent misogyny of AI portraits – Amelia Earhart rendered naked on a bed. *The Guardian*. <https://www.theguardian.com/us-news/2022/dec/09/lensa-ai-portraits-misogyny>
- Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2), 20539517211035955. <https://doi.org/10.1177/20539517211035955>
- Dihal K. (2024, May 11). Provocations & Sharing Concepts in AI & Culture [Session in *Global AI Cultures Workshop*]. ICLR 2024, Vienna, Austria [Online]. <https://iclr.cc/virtual/2024/workshop/20576>
- Dodds, T., De Vreese, C., Helberger, N., Resendez, V., & Seipp, T. (2023). Popularity-driven Metrics: Audience Analytics and Shifting Opinion Power to Digital Platforms. *Journalism Studies*, 24(3), 403–421. <https://doi.org/10.1080/1461670X.2023.2167104>
- Eapen, T. T., Finkenshtadt, D. J., Folk, J., & Venkataswamy, L. (2023, July 1). How Generative AI Can Augment Human Creativity. *Harvard Business Review*. <https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity>
- Equity. (n.d.). *Stop AI Stealing the Show*. Equity. Retrieved 26 February 2024, from <https://www.equity.org.uk/campaigns-policy/stop-ai-stealing-the-show>
- Escalante-De Mattei, S. (2022, December 5). Careful—Lensa Is Using Your Photos to Train Their AI. *ARTnews*. <https://www.artnews.com/art-news/news/does-lensa-ai-use-your-face-data-for-selfies-1234649204/>
- Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Article 1. <https://doi.org/10.1609/aaai.v31i1.10635>

- Flux, E. (2022, December 23). What does the rise of AI mean for the future of art? *The Age*. <https://www.theage.com.au/culture/art-and-design/what-does-the-rise-of-ai-mean-for-the-future-of-art-20221221-p5c81c.html>
- Gaskins, N. (2022, December 17). AI & Techno-Vernacular Creativity. *Medium*. <https://nettricegaskins.medium.com/ai-techno-vernacular-creativity-314281e1be68>
- Gillespie, T. (2024). Generative AI and the politics of visibility. *Big Data & Society*, 11(2), 20539517241252131. <https://doi.org/10.1177/20539517241252131>
- Griffin, A. (2024, February 22). Google takes AI image generator offline over racially diverse historical images. *The Independent*. <https://www.independent.co.uk/tech/google-gemini-ai-images-racially-diverse-b2500730.html>
- Haigney, S. (2023, January 11). AI's Best Trick Yet Is Showering Us With Attention. *The New York Times*. <https://www.nytimes.com/2023/01/11/magazine/ai-photo-filters-selfies.html>
- Hatmaker, T. (2022, December 5). Lensa AI, the app making 'magic avatars,' raises red flags for artists. *TechCrunch*. <https://techcrunch.com/2022/12/05/lensa-ai-app-store-magic-avatars-artists/>
- Heikkilä, M. (2022a, September 16). *This artist is dominating AI-generated art. And he's not happy about it*. MIT Technology Review. <https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/>
- Heikkilä, M. (2022b, December 12). The viral AI avatar app Lensa undressed me—Without my consent. *MIT Technology Review*. <https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/>
- Hill, L. W. (2022, June 12). I tried the AI photo app everyone was using. As someone with a history of disordered body image, it was very triggering to see the results. *Insider*. <https://www.insider.com/woman-with-disordered-body-image-tried-lensa-app-2022-12>
- Hundt, A., Agnew, W., Zeng, V., Kacianka, S., & Gombolay, M. (2022). Robots Enact Malignant Stereotypes. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 743–756. <https://doi.org/10.1145/3531146.3533138>
- Hunter, T. (2023, April 14). AI selfies—And their critics—Are taking the internet by storm. *Washington Post*. <https://www.washingtonpost.com/technology/2022/12/08/lensa-ai-portraits/>
- Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, 51(8), 56–59. <https://doi.org/10.1109/MC.2018.3191268>
- Kamps, H. J. (2022, December 6). It's way too easy to trick Lensa AI into making NSFW images. *TechCrunch*. <https://techcrunch.com/2022/12/06/lensa-goes-nsfw/>
- Keller, S. M. (n.d.). AI Art Gives Users Gender Euphoria—But It's Not Without Controversy. *Out*. Retrieved 4 November 2023, from <https://www.out.com/art/2022/12/07/ai-art-gives-users-gender-euphoria-its-not-without-controversy>
- Kelly, C. (2022, December 11). Australian artists accuse popular AI imaging app of stealing content, call for stricter copyright laws. *The Guardian*. <https://www.theguardian.com/australia-news/2022/dec/12/australian-artists-accuse-popular-ai-imaging-app-of-stealing-content-call-for-stricter-copyright-laws>
- Khan, M. (2023, October 5). People are posting AI-generated yearbook pictures with this viral app. *CNBC*. <https://www.cnn.com/2023/10/05/how-to-create-ai-yearbook-pictures-with-viral-epik-app.html>
- King, M. (2022). Harmful biases in artificial intelligence. *The Lancet Psychiatry*, 9(11), e48. [https://doi.org/10.1016/S2215-0366\(22\)00312-1](https://doi.org/10.1016/S2215-0366(22)00312-1)
- Kircher, M. M., & Holtermann, C. (2022, December 8). How Is Everyone Making Those A.I. Selfies? *The New York Times*. <https://www.nytimes.com/2022/12/07/style/lensa-ai-selfies.html>
- Klee, M. (2022, December 12). A Psychologist Explains Why Your 'Hot AI Selfies' Might Make You Feel Worse. *Rolling Stone*. <https://www.rollingstone.com/culture/culture-features/lensa-app-hot-ai-selfie-self-esteem-1234644965>
- Koidl, K., & Kapanova, K. G. (2020). Introduction to the Special Issue: Re-Imagining a More Trustworthy Social Media Future. *Social Media + Society*, 6(2), 2056305120923007. <https://doi.org/10.1177/2056305120923007>
- Leonardi, P. M., & Jackson, M. H. (2004). Technological determinism and discursive closure in organizational mergers. *Journal of Organizational Change Management*, 17(6), 615–631. <https://doi.org/10.1108/09534810410564587>
- Lovejoy, B. (2023, February 15). AI chat may be big news, but it seems we're over AI photo apps. *9to5Mac*. <https://9to5mac.com/2023/02/15/ai-chat-ai-photo/>
- Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). *Stable Bias: Analyzing Societal Representations in Diffusion Models* (arXiv:2303.11408). arXiv. <https://doi.org/10.48550/arXiv.2303.11408>
- Maimann, K. (2022, December 31). This image is raising money for a Toronto charity. The only problem? It's not real. *Toronto Star*. https://www.thestar.com/news/canada/this-image-is-raising-money-for-a-toronto-charity-the-only-problem-it-is-not/article_9ebc80eb-27ff-5e79-a6bd-acca4beb500f.html
- Marcus, J. (2022, December 10). Artists decry use of AI art: 'I'm concerned for the future of human creativity'. *The Independent*. <https://www.independent.co.uk/news/world/americas/ai-art-lensa-magic-avatar-b2242891.html>
- Markham, A. (2021). The limits of the imaginary: Challenges to intervening in future speculations of memory, data, and algorithms. *New Media & Society*, 23(2), 382–405. <https://doi.org/10.1177/1461444820929322>

- Marres, N. (2015). Why Map Issues? On Controversy Analysis as a Digital Method. *Science, Technology, & Human Values*, 40(5), 655–686. <https://doi.org/10.1177/0162243915574602>
- Martineau, K. (2021, February 9). *What is generative AI?* IBM Research Blog. <https://research.ibm.com/blog/what-is-generative-AI>
- Marx, L. (2010). Technology: The Emergence of a Hazardous Concept. *Technology and Culture*, 51(3), 561–577.
- McCluskey, M. (2022, December 14). Why It's So Hard to Resist Turning Your Selfies Into Lensa AI Art. *Time*. <https://time.com/6240648/lensa-ai-psychology-behind/>
- McKelvey, F., Redden, J., Roberge, J., and Stark, L. (2023, May 24). *CFP: Special issue on (un)Stable Diffusions - Machine Agencies*. <https://machineagencies.milieux.ca/jdsr-cfp/>
- Meyerowitz, A. (2022, December 9). Megan Fox complains that her AI profile photos have been sexualised by viral app. *Glamour UK*. <https://www.glamourmagazine.co.uk/article/megan-fox-lensa>
- Miltner, K. M., & Baym, N. K. (2015). The Selfie of the Year of the Selfie: Reflections on a Media Scandal. *International Journal of Communication*, 9, 1701–1715.
- Miltner, K. M., & Highfield, T. (2024). *The Possibilities of 'Good' Generative AI in the Cultural and Creative Industries* (p. 10). The British Academy. <https://www.thebritishacademy.ac.uk/publications/the-possibilities-of-good-generative-ai-in-the-cultural-and-creative-industries/>
- Natale, S., & Ballatore, A. (2020). Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence*, 26(1), 3–18. <https://doi.org/10.1177/1354856517715164>
- Nicoletti, L., & Bass, D. (2023, June 7). Humans Are Biased. Generative AI Is Even Worse. *Bloomberg*. <https://www.bloomberg.com/graphics/2023-generative-ai-bias/>
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.
- O'Brien, S. A. (2022, December 11). 'Magic' AI Avatars Are Already Losing Their Charm. *Wall Street Journal*. <https://www.wsj.com/articles/lensa-ai-avatars-women-11670705062>
- Offert, F., & Phan, T. (2022). *A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space* (arXiv:2211.06323). arXiv. <https://doi.org/10.48550/arXiv.2211.06323>
- Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY*, 35(4), 927–936. <https://doi.org/10.1007/s00146-020-00965-5>
- Ovide, S. (2022, December 10). Your selfies are helping AI learn. You did not consent to this. *Washington Post*. <https://www.washingtonpost.com/technology/2022/12/09/chatgpt-lensa-ai-ethics/>
- Perez, S. (2023, February 13). The AI photo app trend has already fizzled, new data shows. *TechCrunch*. <https://techcrunch.com/2023/02/13/the-ai-photo-app-trend-has-already-fizzled-new-data-shows/>
- Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, 14(3), 399–441. <https://doi.org/10.1177/030631284014003004>
- Portwood-Stacer, L. (2013). Media refusal and conspicuous non-consumption: The performative and political dimensions of Facebook abstention. *New Media & Society*, 15(7), 1041–1057. <https://doi.org/10.1177/1461444812465139>
- Prisma Labs, Inc. (n.d.). *Lensa's Magic Avatars Explained*. Notion. Retrieved 22 November 2023, from <https://prismalabs.notion.site/prismalabs/Lensa-s-Magic-Avatars-Explained-c08c3c34f75a42518b8621cc89fd3d3f>
- Rafeeq, A., & Jiang, S. (2018). From the Big Three to elite news sources: A shift in international news flow in three online newspapers TheNational.ae, Nst.com.my, and Nzherald.co.nz. *The Journal of International Communication*, 24(1), 96–114. <https://doi.org/10.1080/13216597.2018.1444663>
- Rekhi, D. (2022, December 17). *Cyber experts train lens on Lensa-like apps—ET Telecom*. ETTelecom.Com. <https://telecom.economictimes.indiatimes.com/news/cyber-experts-train-lens-on-lensa-like-apps/96292115>
- Roberts, M. (2022, May 12). The Euphoric Highs & Problematic Lows Of AI Avatar Art. *Refinery20*=9. <https://www.refinery29.com/en-gb/ai-avatar-art>
- Roose, K. (2022, October 21). A.I.-Generated Art Is Already Transforming Creative Work. *The New York Times*. <https://www.nytimes.com/2022/10/21/technology/ai-generated-art-jobs-dall-e-2.html>
- Sandoval-Martin, T., & Martínez-Sanzo, E. (2024). Perpetuation of Gender Bias in Visual Representation of Professions in the Generative AI Tools DALL·E and Bing Image Creator. *Social Sciences*, 13(5), Article 5. <https://doi.org/10.3390/socsci13050250>
- Santos, K. M. L. (2020). The bitches of Boys Love comics: The pornographic response of Japan's rotten women. *Porn Studies*, 7(3), 279–290. <https://doi.org/10.1080/23268743.2020.1726204>
- Scherer, M. (2024, January 4). The SAG-AFTRA Strike is Over, But the AI Fight in Hollywood is Just Beginning. *Center for Democracy and Technology*. <https://cdt.org/insights/the-sag-aftra-strike-is-over-but-the-ai-fight-in-hollywood-is-just-beginning/>
- Scheufele, D. A., & Lewenstein, B. V. (2005). The Public and Nanotechnology: How Citizens Make Sense of Emerging Technologies. *Journal of Nanoparticle Research*, 7(6), 659–667. <https://doi.org/10.1007/s11051-005-7526-2>
- Senft, T. M., & Baym, N. K. (2015). Selfies Introduction ~ What Does the Selfie Say? Investigating a Global Phenomenon. *International Journal of Communication*, 9(0), Article 0.

- Shane, T.S. (2023). AI incidents and ‘networked trouble’: The case for a research agenda. *Big Data & Society*, 10(2), 20539517231215360. <https://doi.org/10.1177/20539517231215360>
- Silberling, A. (2022, December 1). Lensa AI climbs the App Store charts as its ‘magic avatars’ go viral. *TechCrunch*. <https://techcrunch.com/2022/12/01/lensa-ai-climbs-the-app-store-charts-as-its-magic-avatars-go-viral/>
- Smith, A. (2022, December 20). Lensa AI image app helps some trans people to embrace themselves | Context. *Openly News*. <https://www.openlynews.com/i/?id=0d0b52f6-0fa4-4c27-9881-f689ad9fbe87>
- Snow, O. (n.d.). ‘Magic Avatar’ App Lensa Generated Nudes From My Childhood Photos. *WIRED*. Retrieved 29 November 2023, from <https://www.WIRED.com/story/lensa-artificial-intelligence-csem/>
- Sobande, F., Fearfull, A., & Brownlie, D. (2020). Resisting media marginalisation: Black women’s digital content and collectivity. *Consumption Markets & Culture*, 23(5), 413–428. <https://doi.org/10.1080/10253866.2019.1571491>
- Stability AI. (2022, August 22). *Stable Diffusion Public Release*. Stability AI. <https://stability.ai/news/stable-diffusion-public-release>
- Sun, M., & Harmon, S. (2024, March 6). ‘The worst AI-generated artwork we’ve seen’: Queensland Symphony Orchestra’s Facebook ad fail. *The Guardian*. <https://www.theguardian.com/culture/2024/mar/06/queensland-symphony-orchestra-ai-facebook-ad-criticism>
- Sung, M. (2022, December 7). Lensa, the AI portrait app, has soared in popularity. But many artists question the ethics of AI art. *NBC News*. <https://www.nbcnews.com/tech/internet/lensa-ai-artist-controversy-ethics-privacy-rcna60242>
- Thiel, D. (2023, December 20). Investigation Finds AI Image Generation Models Trained on Child Abuse. *Stanford Cyber Policy Center*. <https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse>
- Tiidenberg, K. (2018). *Selfies: Why We Love (and Hate) Them*. Emerald Publishing.
- Tran, T. H. (2022, December 10). Image Apps Like Lensa AI Are Sweeping the Internet—And Stealing From Artists. *The Daily Beast*. <https://www.thedailybeast.com/how-lensa-ai-and-image-generators-steal-from-artists>
- Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, 9(11), 40–53. <https://doi.org/10.22215/timreview/1282>
- Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022). Limits and Possibilities for “Ethical AI” in Open Source: A Study of Deepfakes. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2035–2046. <https://doi.org/10.1145/3531146.3533779>

Automated hiring platforms as mythmaking machines and their symbolic economy

Etienne Grenier and Nicholas Chartier-Edwards

Institut national de la recherche scientifique, Canada

✉ etienne.grenier@inrs.ca

Abstract

Automated hiring platforms offer critical Artificial Intelligence (AI) researchers a privileged site for the study of technological controversies and their unfolding as they often end up entangled in scandals. In 2019, an official complaint filed with the Federal Trade Commission (FTC) outlines the main controversy created by such systems, directly targeting HireVue's platform. According to the FTC, these systems "evaluate a job applicant's qualifications based upon their appearance by means of an opaque, proprietary algorithm" (FTC, 2019, p.1). Existing critical literature outlines how Basic Emotions theory (Ekman, 1999) and its deployment through an AI-powered system fuel ventures such as automated hiring platforms (Stark, 2018). Therein lies a form of opportunism where the use of available audiovisual data and the revalorization of heavily criticized and simplistic theories about the human mind create discriminatory automated decision-making grounded in bogus science (Crawford, 2021). Even if HireVue slightly modified its product by removing the video analysis component in reaction to the critiques formulated by regulatory bodies, the company kept AI-powered voice analysis, its one-sided video interview technique and the multitude of gaming assignments designed to evaluate candidates. Going beyond scandals and hype, this article aims to tackle the inevitable political economy generated by automated hiring systems. Indeed, job seekers are confronted with a system that disrupts a well-established, sociologically stable technique, the interview, and its technical object counterpart, the resume. We assert that the gaps of nonknowledge (Beer, 2023) created by the one-sided interviews break the commonly accepted interactionist framework that originally informed the hiring process. Using HireVue as a case study, we lay bare the political economy of AI-powered automated hiring platforms, highlighting how individuals assess their capacity to acquire agency in the face of opaque technologies.

Automated Hiring; HireVue; Reddit; Nonknowledge; Political Economy

1. Introduction

On January 1, 2020, the State of Illinois passed the Artificial Intelligence Video Interview Act (AIVI) which regulates the use of automated hiring platforms (AHPs) based on AI components. More specifically, section 5.2 of the act states that job applicants screened by such systems must be provided

with information “explaining how the artificial intelligence works and what general types of characteristics it uses”¹ in its evaluation before the interview. Since AHPs “typically use natural language processing (NLP), computer vision, or supervised machine learning techniques that claim to predict job performance” (Sloane et al., 2022, p.1), and that such systems are known for their opacity, the application of legislation such as the AIVI act will prove to be difficult. The controversial nature of volatile, black-boxed systems (Macospol, 2007; Venturini, 2010) extends well beyond their technical part to engulf a wide variety of social actors entangled in a mesh of relations taking the shape of conflicts, negotiations, and strategic interactions.

Of late, a specific AHP vendor piqued the interest of privacy advocate groups and became the subject of frequent coverage in the general press: HireVue (Crawford, 2021; Doctorow, 2021; Knight, 2021). The company’s product, an AI-powered platform aiming for the complete automation of the hiring process, came to be targeted by an official complaint filed to the Federal Trade Commission (FTC) by the Electronic Privacy Information Center (EPIC) in 2019. According to EPIC, these systems: “evaluate a job applicant’s qualifications based upon their appearance by means of an opaque, proprietary algorithm.” (FTC, 2019, p.1)². The extended deployment of this specific platform,³ following its enthusiastic promotion by HR consulting firms and counselors, created outcry amongst job applicants who flocked to online forums and fueled a public discussion about the effects of such technology. Reddit, precisely the *r/recruitinghell* subreddit,⁴ has become a site of intense debate where users rant about their experience involving HireVue and, furthermore, discuss potential strategies that could be adopted to win the automated job-interview game, might they be technical hacks, *mise-en-scène* tricks or behavioral adaptation techniques.

The aim of this article is to investigate the symbolic economy created by AHPs and to contribute to the growing field of critical AI studies (Roberge & Castelle, 2021). When discussing the recursion at work in societies where AI-laced technologies have been deployed and operating for some time, David Beer claims it is “built upon feedback loops, multiple feedback loops, each endlessly feeding into the next” (Beer, 2022, p. 1). It is through the study of the controversiality of HireVue and its structuring effect on online communities that we identify a grounded example of these recursive social forms. To do so, we performed a cross analysis of HireVue’s corporate content with a non-participant observation of various online forums where the company’s AHP was discussed. The technological disruption of the hiring interview, a well-established gatekeeping social ritual, combined with the scarcity of contextual information resulting from the opacity of the AHP, creates an environmental pressure that shapes job applicants’ online communities. Beyond the despair and angst expressed on Reddit, our research led us to witness the emergence of a symbolic economy situated around the black box, where job applicants would trade dubious knowledge concerning the AHP in the hope of fostering agency amongst the community or on a personal level. By studying online exchanges on forums and corporate communications, we observed that pressure exerted by a specific sociotechnical system, HireVue’s AHP, over a vastly distributed human group, the job applicants, has a structuring effect on their loosely coordinated online community. These communities spontaneously form and their members engage to correct the knowledge gaps situated at the root of the disruption caused by the AHP in the social ritual of interviews. We can witness within online forums that the applicants as well as HR experts participate in a parallel economy where the exchanged currency is constituted of dubious information, hoping to acquire agency against a faceless gatekeeper managing the social selection process. Looped around

¹ State of Illinois, Public Act 101-0260, 2020. <https://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=101-0260>

² <https://epic.org/documents/in-re-hirevue/>

³ Boeing, GE, Unilever, Deloitte, PwC, Mercedes-Benz, Kraft Heinz, Delta, Hilton and St. Jude Children’s Research Hospital, amongst many others.

⁴ In the context of our data collection, of the 132 Reddit posts mentioning HireVue, 109 were located in the *r/recruitinghell* subreddit.

a technological object, social actors engage in trade and circulate, changing roles as their status morph within this circuit.

2. AHPs & critical AI studies

Our study of AHPs began with a review of some of the existing literature from the broad field of critical AI studies which “trains itself on the entire sociotechnical and political process of modern machine learning from genesis to impact and back again” (Roberge & Castelle, 2021, p. 3), focusing on the deployment of automated hiring platforms. Our first assessment showed that the subject of automated hiring was already treated in corpuses concerned with bias making by algorithmic systems rooted in junk science. Scholars such as Luke Stark, Mona Sloane, Jevan Hutson, Kate Crawford have all contributed to this body of literature, notably by demonstrating how automated hiring, which often mobilizes facial recognition (FRT) and voice recognition technologies (VRT), reactivated old debunked pseudo-sciences such as physiognomy and phrenology (Stark & Hutson, 2021; Crawford, 2021; Sloane et al., 2022). The epistemological roots of the psycho-computational complex (Stark, 2020) can be found in the work of 19th century British scientist Sir Frances Galton and still rely heavily on the use of visual cues tied to bodily signs (Stark, 2018). To support the wild claim “that they reveal the true potential of job candidates” (Sloane et al., 2022, p. 1), current AHP systems pair up FRT to a truncated understanding of Basic Emotion Theory (BET). According to Stark, this clunky assemblage does “not lend sufficient scientific evidence for even nuanced arguments grounded in the Motivational tradition,⁵ much less the often sensationalist claims made by many FRT providers.” (Stark & Hoey, 2021, p. 790) While the recursion of pseudoscience is not uninteresting to us, we turned our attention to the question of power within and without AI systems, mainly the precise workings of how each system operates in its own private, autonomous, micro-political economy (Farrell & Fourcade, 2023). Rahman recently described opaque third-party algorithmic evaluations as constitutive elements of “‘post-panoptic’ environment for workers” which the author dubbed the invisible cage (2021, p. 949). Acting as an algorithmic evolution of the iron cage theorized by Weber, these environments trigger reactivity amongst workers as “they have difficulty identifying the evaluations’ underlying expectations, including what is rewarded or punished.” (Rahman, 2021, p. 977) Beyond stating they are resisting such evaluations through only platforms dedicated to knowledge sharing, where they can “help each other learn new systems and practices” (Kellogs et al., 2020, p. 392), we wondered how such knowledge exchanges would be organized.

By shifting our focus from the study of “algorithmic harm” and “algorithmic resistance,” two already explored approaches described above, to the study of an AI-enabled symbolic political economy, we aim to sketch out the understanding of how a multitude of systems, operating in their own realms while potentially overlapping, reconfigures our societies through recursive mechanisms. This article is aligned with the approaches of scholars that examine the reconfiguration of power relations by AI, such as Farrell and Fourcade’s work on the creation of private knowledge regimes by chatbots (Farrell & Fourcade, 2023), Beer’s theoretical writings on the social recursiveness emerging around AI systems (Beer, 2022; 2023) and, finally, the work of Roberge and affiliates such as McKelvey on the becoming of AIs black boxes opacity as a mode of governance which prompts a form of recursive power (Roberge & al., 2020; Roberge & McKelvey, 2023).

⁵ BET is an offshoot of the Motivational psychological tradition in the study of affect. Its founder notoriously claimed that “the primary motivational system is the affective system.” (Tomkins, 2008, p.4) Paul Ekman subsequently developed BET claiming there exists six universal basic emotions which are tied to specific facial expressions. (Ekman, 1992)

3. Methodology

The methods we chose to study HireVue's AHP are highly inspired by the work of science and technology studies (STS) scholars such as sociologist Angèle Christin and anthropologist Nick Seaver regarding the challenges posed to social sciences by the opacity of the algorithmic black box, which often incarnates either corporate secrecy, mathematical complexity, or pure unexplainability when in the presence of neural net-based technologies (Christin, 2020; Seaver, 2017). For critical AI studies authors, the algorithmic opacity is not in itself a hurdle; the conjugation of classical social science methods and new digital methods, or computational science knowledge, can inform us on the inner workings of these technologies. As Seaver argues, algorithms and individuals mutually influence themselves, meaning that algorithms can be known through the ways they influence society (Seaver, 2017). Going further, Seaver urges social scientists to stop feeding into the "algorithmic drama" of the black box by redirecting their gaze to the social and material impacts of automated technologies (Seaver, 2022). This opens the perspective of studying automated technologies from the *exterior*, rather than fixating on the *interior* of the algorithmic black box. Social scientists can work with exterior materials such as interviews with individuals, computer science texts, mathematical models, corporate narratives, lab ethnography, financial reports and so on. This methodological approach is what Christin names the "algorithmic refraction," which she defines in the following way:

The concept of refraction is derived from physics, where it refers to the changes in direction and strength that occur whenever a wave of light or sound passes from one medium to the next. Applied to algorithms, studying refraction entails paying close attention to the changes that take place whenever algorithmic systems unfold in existing social contexts—when they are built, when they diffuse, and when they are used. (Christin, 2020, p. 906)

To gain "ethnographic insight into how job applicants experience AHPs" (Sloane et al., 2022, p.7), we decided to engage in non-participant observation on forums where individuals would express themselves on automated hiring, whether expressing concerns, criticizing the technology, or expressing how an AI-piloted interview made them feel, as well as individuals praising HireVue. We proceeded with the data collection⁶ for this study in February 2023⁷, gathering online-available material, mainly testimonies from the Reddit subreddit dubbed *r/recruitinghell* from individuals who claimed they passed a HireVue interview⁸, but also videos posted on YouTube from individuals claiming to be human resources specialists⁹ peddling and selling "counsels" and "advice" and even, in some cases, AI-powered technology for interview practice.

Following a "thick data" approach in order "to reduce the breadth of data (the number of datapoints) while enhancing their depth (i.e. 'thick-ness' of each datapoint)" (Latzko-Thot et al., 2022, p. 202), we developed a process to collect and categorize posts from Reddit. This specific platform was chosen since its "masked" commenters are "often discussing lengthy and detailed personal experiences, sometimes asking for help, support and information, and sometimes simply sharing their thoughts and opinions." (Adams, 2024, p. 48) The expression of negativity on this platform, bolstered by anonymity, could be revealing of the harm caused by HireVue's AHP. Multiple qualitative studies (Low et al., 2020; Slemon et al., 2021) powered by natural language processing (NLP) and enabled by the Pushshift dataset (Baumgartner et al., 2020) gave insights to the research communities about the expressive content available on the platform. As an example, studies focusing on eating disorders conducted on the same platform revealed that "feelings of frustration, anxiety and loneliness were shown to increase in presence within linguistic discussions by Reddit users." (Adams, 2024, p. 50) We concluded that harmful experience involving HireVue's AHP could be narrated in lengthy detail by Reddit users in a non-chastised manner, hence the choice of said platform for our inquiries.

⁶ A digital archive was created from the collected textual content for further analysis.

⁷ Publications on Reddit and YouTube mentioning HireVue were exhaustively collected at the date.

⁸ Reddit: 139 posts amongst which 131 were located on the *r/recruitinghell* subreddit ; YouTube: 16 posts.

⁹ YouTube: 16 posts.

The first phase of data collection was conducted by manually scraping content from Reddit using a simple search term: “HireVue”. We exhaustively recuperated all posts (139) where the term was occurring. Quite surprisingly, the vast majority of them (131, which corresponds to 94.25% of the whole) were located on the r/recruitinghell subreddit. The fact that this particular subreddit, being mainly dedicated to the shared misadventure narratives of job applicants, was already an indicator that its posts would feature “liberated voices” and would be in that sense quite critical of AHPs. Most posts located on other subreddits¹⁰ were from job applicants asking about a future Hirevue interview for highly skilled positions (accountants and engineers) in international firms.

The second phase of data collection was aimed at the YouTube platform. Often used as marketing tools to advertise services, professional YouTube accounts could lead us to a type of content concerning HireVue that would differ from the disgruntled job applicant comments on Reddit, therefore acting as a counterpoint and adding depth to the object of study. The platform was selected as a site of inquiry as it is used by self-proclaimed experts to peddle some basic counseling as a kind of fishing practice to then sell more “profound” advice. In order to recuperate the so-called experts’ discourses on the platform, we used the following search terms: “HireVue”, “HireVue Experts”, “Tips for HireVue interview.” The collected discourses¹¹ of the experts were transcribed for further analysis.

During the third phase of data collection, we drew material from HireVue’s website to understand the corporate narrative and its promises made to the potential buyers of the technology but also, potential individuals who would undergo HireVue interviews (Borup et al., 2006). Additionally, we collected the very few interactions HireVue’s representative had with job applicants on Reddit¹². The triangulation of collected materials such as online testimonies, counsel peddling and corporate narrative, which we will couple with already existing critical AI studies theory on the 21st century AI paradigm of machine learning and neural nets (Roberge & Castelle, 2021; Beer, 2023), sets the foundation of our analysis. While the controversial nature of Hirevue’s AHP is made apparent by the vast majority of posts we collected from Reddit,¹³ we firmly believe that the best approach to explore this controversy does not lie in situational analytics (Marres, 2020) or computational methods but rather in a more traditional qualitative interpretative approach where the discourse registers would emerge as well as the entirely speculative nature of the information exchanged on the different forums.

4. Case study: HireVue and its publics

The following section recapitulates our findings at a granular level.¹⁴ This is where the principal data upon which our argument is built are explained. Large-scale technological deployment, shaky knowledge, one-sided interactions, alienation, disengagement, shared ruses, and various adaptation strategies, both corporate and individual, constitute the landscape which prompts the emergence of a symbolic and yet deeply political economic circuit. Through the testimonies left by actors enmeshed in controversial

¹⁰ /Big4, /Boeing, /HireVue_, /deloitte, /FinancialCareers, /recruiting.

¹¹ Saturation was rapidly reached with these contents as their messages were similar both in form and intent.

¹² Reddit: 4 posts.

¹³ The unfavorable press coverage and the recent legislative action mentioned in the introduction of the article are also clear controversy markers. As Callon states, controversies, especially technological controversies, are negotiated both through scientific publications (the forum constituant) and through popular media (the forum officieux). Actors can occupy either or both of these positions and the interactions between the forum constituant and forum officieux help define the overarching narrative of the controversy (Callon, 2006). We thus hoped to bring together the discontent expressed on the r/recruitmentell subreddit with the sociological literature criticising the phenomenon of automated hiring platforms.

¹⁴ In order to protect the anonymity of the commenters, we chose not to disclose their usernames. We also limited the quantity of quoted material to further protect their identity. All of the few quotes that were incorporated in this article have been tested with reverse search to make sure that it would prove difficult to trace them back to their original context. These measures have been taken in order to address issues about the ethical use of scraped Reddit data that were raised in recent literature (Adams, 2024).

relationships centered on HireVue's AHP distinctly appear the pressure-applying mechanics that shape said circuit.

4.1 Key findings: HireVue's corporate communications

4.1.1 Scalability, seamlessness & speed

HireVue presents its AHP as a universal tool that can be deployed in a wide variety of contexts. It can scale efficiently to tackle the needs of large scale or medium-size businesses. The AHP is also presented as a completely integrated solution that “makes the hiring process seamless.” One of the main selling points of the platform is the speed at which it will treat many job applications. According to the corporate website, “Hiring teams are facing a competitive market, and the first-to-offer has the first pick of talent.” (HireVue, 2023). In the context of a labor shortage crisis, this argument might be appealing to potential customers.

4.1.2 Adaptation to critique

Following the many critiques targeted at AHP, HireVue deployed a contrarian discourse where the AHP could be used to “support your diversity goals,” instead of being tagged as a discriminatory device. Also, following the change in regulations, lawsuits, and public outcry, HireVue dropped the AI-powered video analysis component from its platform.¹⁵ A vague statement on the corporate websites states HireVue's wish, “to evolve as we work with our customers, jobseekers, technology partners, ethicists, legal advisors, and society at large.” Meanwhile, a class action was filed against HireVue on January 27th 2022 for multiple violations of the Illinois Biometric Information Privacy Act (Willner & Sabna, 2022).

4.1.3 Short-lived engagement on public forums

An official HireVue representative answered a total of four posts made by job applicants. Mention of “24/7 support team” able to answer email or phone calls was made. If, initially, support was offered,¹⁶ it seems HireVue's PR representative has subsequently vanished from Reddit.

4.2 Key findings: HR experts & counselors

4.2.1 AI or not?

While HireVue claims to have eliminated the AI video analysis component of their platform, most HR experts capitalize on the fear created by such techniques. “I know it's a bit disturbing that there's an algorithm essentially analyzing you, but try not to think about it,” suggests one of these experts on YouTube. According to these specialists, AHPs can tell whether a candidate is nervous or not as well as confident or not. Eye contact, body language and intonation are mentioned as cues being analyzed during the interview. Moreso, the transcribed content of what an applicant said during an interview is compared to a list of desired keywords predetermined by the potential employer.

4.2.2 Ignorant counselors

It is very difficult to tell if any of the counselors offering advice about HireVue have any real, direct experience with the platform. Some explicitly expose they do not know how the AHP operates but since they have extensive experience as interview coaches, they should be knowledgeable. Most of the advice given out to eventual applicants is generic and claims about the AHP are entirely speculative.

¹⁵ Hirevue's CEO Kevin Parker claims that AI-powered video analysis “was adding some value to customers, but it wasn't worth the concern” (Knight, 2021, para. 5).

¹⁶ “So if things go wrong in your interview, or if you need a webcam, tips and tricks, someone to talk to or anything like that, we are here,” wrote a HireVue representative.

4.2.3 *Technical advice, mise-en-scène and behavioral adaptation*

The experts suggest investing in decent audiovisual equipment. A semi-pro microphone, camera and lighting kit should drastically improve the technical quality of the recording. It was also suggested to test the equipment prior to the interview. Many experts insist on being mindful of the surroundings and create a decent backdrop for the interview. Besides these counsels, applicants should spend time getting “more comfortable on camera” and train themselves. According to the experts, “watching yourself back and noting what you’re doing wrong” is part of the preparation for an AHP interview.

4.3 *Key findings: Job applicants’ forum posts*

4.3.1 *What really is on the record (and why)?*

The commenters disagree whether the video is recorded continuously, even when the participant is not submitting a video answer to a question. A few applicants complained that the recorded videos could be kept forever. Some participants claim that they could be used by HireVue as training data for undisclosed ulterior motives. There is also a debate concerning whether the videos are viewed at all by human recruiters. Some commenters claim the videos are treated automatically and that the scores gained through game-like assessments are what really count.

4.3.2 *Dehumanizing and dystopian*

Automated interviews are perceived by many commenters as disrespectful and dehumanizing. Some applicants claim that AHPs are in fact discriminatory and that HireVue “furthers the transition from workers people being just numbers [sic].” Even one participant claims that the system comes straight out of cyberpunk fiction.

4.3.3 *One-sidedness*

HireVue, like many AHPs, uses an interview model where job applicants are left alone in front of their webcam. As they do not have counterparts to talk to, some commenters felt that they have undergone a disrespectful process that is not a “real” interview. The one-sidedness characterizing HireVue’s platform is deemed unfair by many participants since: “Interviewing is a two-way street and we candidates have just as much right to ask questions of them as they have a right to ask them of us.” Furthermore, a commenter voiced a concern many applicants have about this characteristic of AHPs: “The one-sidedness of it means that it inherently is bad for getting a good picture of a potential candidate.” According to many commenters, it appears to be very difficult to create a good impression in the absence of an interlocutor who could give direct feedback through verbal and non-verbal cues.

4.3.4 *Lack of adaptation and effort*

Commenters who experienced HireVue’s AHP told forum members that some questions were being used cyclically from one interview to the other. Also, according to many applicants, some of the behavioral questions of the assessment are so generic they do not relate in any way to the job for which they made an application. Some interviews feature coding quizzes, which were deemed badly thought out by some applicants. A commenter who had previous experience as a recruiter even declared about the AHP: “This is one of those things that sound good in theory but I haven’t come across a company to use something like this for more than two years.” Another blamed the laziness of the recruiters who use these AHPs: “Recruiters don’t actually put any effort in, so the AI does all the work.”

4.3.5 *Shared ruses: Technique, behavior and mise-en-scène*

Some participants provide technical hacks to unlock the interview questions from the web browser in advance. There even exists GitHub repositories that contain code to perform this type of exploit. Also, participants share tricks involving app minimization to “stall the timer.” Other suggestions by applicants involved changing the way one expresses themselves, “you gotta over exaggerate the way you explain

things,” and make sure to sound “super excited, super enthusiastic.” According to a commenter, the AI analysis part of HireVue’s AHP will favor “people with more clear-cut facial expressions.” Extensive discussions about clothing and more specifically, its specific color, could be found in many threads.

5. Data analysis

5.1 *An ongoing controversy and job applicants’ online communities*

Contradictions between corporate storytelling, scientific literature, and user experience characterize HireVue’s AHP. What is this technological platform exactly measuring? How are its AI components used and does it discriminate unfairly against specific groups of users? Online, such questions are the subjects of heated debates. The public discussion happening on Reddit pits the numerous negative comments posted by job applicants against the corporate claims of fairness and equity. This apparently face-to-face confrontation is complexified by the presence of seemingly independent agents, the so-called HR experts, who contribute to the public debate while trying to profit from the situation. While sorting the numerous posts we collected into broadly defined categories,¹⁷ we noticed that the job applicants involved in the controversy could switch roles. For example, one-time job applicants seeking to learn more about the AHP could end up reposting on the same thread after their interview. The same could be said of commenters who could, at first, complain about the degrading nature of the AHP and, in the context of another post, provide directives to other applicants so they increase their chances of success on their next interview. Depending on their level of knowledge, based either on direct experiences of an AHP-conducted interview or on a basic understanding of the hiring process’s unwritten rules, these social actors can modify their status, gain some agency, and become counsel providers themselves. Inversely, they might roll back to being a commenter or, even, a knowledge seeker, if they experienced adverse results following an interview. These shifting roles amount to movements within an emerging circuit anchored on the AHP. The shared online bits of knowledge, even if they are mostly based on imagination since AHPs are highly black boxed,¹⁸ highlights the existence of an informal support system that helps applicants to cope with a difficult situation. As Venturini wrote, “Controversies are the place where the most heterogeneous relationships are formed.” (2010, p. 261) A wide variety of actors from different backgrounds met online and created an exchange system organized around a technical object, HireVue’s AHP, where roles could switch depending on the amount of knowledge one has. By refusing to disseminate knowledge concerning the mode of operation of their AHP or to engage online with job applicants, HireVue¹⁹ created a form of informational scarcity, which creates a power imbalance situated at the level of the job applicant and the machine.

¹⁷ In order to get a grasp of the ongoing public debate, forum posts were tagged with the following properties: actors (HR Specialists, HireVue PR, Job Applicants, Onlookers), roles (Knowledge-seekers, Commenters [Specialized, Amateur or Experimented], Prescribers [Specialized, Amateur or Experimented]), expression register (critique, justification or adaptation) and then, types of critiques (HireVue Criticism, Socio-Economic Criticism, Alienation), types of adaptation (Behavior, Mise-en-scène, Technical ruse, Documentation) and types of justification (Socio-economic, Trickery, Automation, Corporate).

¹⁸ The fact that HireVue rapidly stopped its online engagement with the community only worsened this situation.

¹⁹ An additional argument based on HireVue’s website content could be made to support the existence of a deliberate corporate obfuscation strategy. Pretending to offer a custom-tailored hiring solution, the company claims are debunked by the confirmation that many interview questions are in fact recycled. Also, the website’s vague ethical declaration hinting in favor of cooperation with a wide variety of social actors is contradicted by the ongoing legal actions opposing the company to privacy advocacy and labour rights groups. The narrative deployed by the company contradicts factual evidence about its product and its effects.

5.2 Uneasy interviews: disrupted interactions and dysfunctional strategies

While some meta-analysis of reactions to AHPs present a balanced account, positive against negative, of job applicants' attitudes who submitted to such a process,²⁰ our findings indicate that most commenters on Reddit and YouTube are voicing negative opinions. Indeed, the one-sidedness of the AHP interview generates many criticisms. The traditional social form of the job-seeking process relied on the mediation of a technical object, the resumé, and of a human technique, the interview. By replacing the human recruiter with an AHP, HireVue claims it can now go beyond the resumé and “crack the technical hiring code.” (HireVue, 2019)

Through the concepts of face interaction, performance disruption and impression management, Goffman's microsociological framework of interactionism can help us understand further the distress of interviewees facing a faceless machine. The interaction ritual provides an interesting frame of reference to analyze job interviews since those are established form of interactions implying, “the standardization of bodily and vocal behavior.” (Goffman, 1983, p. 3) Like most social rituals, they are heavily codified and rely on conventions, establishing the “rules of the game.” If job interviews are understood as “people-processing encounters, encounters in which the ‘impression’ subjects make during the interaction affects their life chances,” and if “this processing is ubiquitous; everyone is a gatekeeper in regard to something,” then, what happens if said gatekeeper is faceless? (Goffman, 1983, p. 8) The one-sidedness characterizing AHP interviews breaks the traditional form of a long-established social ritual and creates gaps. The impossibility of a dialogue with a non-existent other, while having to perform as if there was an audience, pushes job applicants in an unknown territory: “The situation may cease to be defined, previous positions may become no longer tenable, and participants may find themselves without a charted course of action.” (Goffman, 1956, p. 155) More than a simple dysfunction, the one-sidedness of AHP interviews plays into the knowledge scarcity scheme already at work. Not only adding to the whole opaqueness of the AHPs black box, it also creates an asymmetric power relation where job applicants have to supply all the necessary visual and auditory cues to support their claim for work. This contradicts the very basic assumptions of an encounter²¹. As one of the applicants said on the forum, “It should go both ways.” AHPs create a situation where the other side is disembodied, acting like a negative presence, sucking in information, and providing none.

This last characteristic of AHP interviews inclines us to classify them into the regime of “synthetic situations” as described by Karin Knorr Cetina. Building on previous research conducted with globally (and algorithmically) connected financial traders,²² the author identifies a new type of situation that imposes a reevaluation of the standard encounter described by Goffman. More than a simple “disorderly interactional arrangement struggling with problems of differential access, orientation and perspective, and coordination,” a synthetic situation “is a composite of information bits that may arise from many areas around the world and feature the most diverse and fragmented content.” (Cetina, 2009, p. 74, p. 69) One of the key features of these synthetic situations²³ is that their existence relies on scopic devices which

²⁰ See Hunkenschroer & Luetge, 2022, for example.

²¹ The classical Goffmanian interaction starts with a bilateral information acquisition process: “When one enters the presence of others, he will want to discover the facts of the situation. Were he to possess this information, he could know, and make allowances for, what will come to happen [...] Full information of this order is rarely available; in its absence, the individual tends to employ substitutes—cues, tests, hints, expressive gestures, status symbols, etc.—as predictive devices.” (Goffman, 1956, p. 155)

²² See Cetina, K. K. & Bruegger, U. (2002). Global Microstructures: The Virtual Societies of Financial Markets. *American Journal of Sociology*. 107(4). 905-950. <https://doi.org/10.1086/341045>.

²³ In line with Cetina's work, we consider our research endeavors to reside entirely in the realm of synthetic situations. The interviews as well as the comments on the platform or the knowledge exchanges implicating the self-proclaimed experts are to be understood as events making sense within a digital context without having to trace back these interactions to real-life incarnated population segments.

act as “instrument[s] for seeing or observing” (2009, p. 64). Beyond allowing the situation participants to coexist in remote locations, the scopic devices “make visible, project, and record things that cannot be seen in the physical situation or can be seen only in an informationally deprived manner—but that are causally implicated in the progress of the situation and its later outcomes.” (2009, p. 82) This definition fits the particulars of technology supporting the AHP, as the participants in such interviews provide not only their audiovisual answers to questions but also a trove of information collected during their course such as their behavior during the game assignments. The main difference between the traders observed by Cetina and the HireVue’s job applicants is that the latter’s experience is akin to a one-way mirror: the scopic device is looking at them, transferring algorithmic power to the eventual recruiters in an asynchronous and asymmetrical relation. In this case, the object to be gazed at and manipulated is not the market but the workers.

Building on the concept of fissures in algorithmic power—“moments in which algorithms do not govern as intended”—we can observe in this specific case that “algorithmic power enables and constrains social action, it entails domination and counteractions, and it is practiced—not possessed.” (Ferrari & Graham, 2021, p. 815, p. 818) While the job applicants seem to be stuck in some asymmetric power struggle with a faceless and ghastly opponent, they still provide an informal response through the forum and create a social network dedicated to the matter. However, the cumulation of job applicants’ attitude toward the AHP opaqueness (even if it is sometimes rooted in ruse), and the trauma caused by one-sided interviews, do not actuate in an organized political response in the name of workers’ rights. If “platforms bring reshaped modes of social contestation into being,” (Ferrari & Graham, 2021, p. 827) they do not aim for revolution or organize aggression against the deployment of AHPs. The protestations expressed on the forum, while showing some level of solidarity, fuels an individualistic market-like attitude where one can beat the game.

6. The political economy of machinic mythmaking: Nonknowledge and agency

To understand the way that HireVue’s AHP modulates individuals’ lives through both behavioral modification and material conditions in the form of access to a job, we turned to the sociological tradition of political economy²⁴. Precisely, we aim to demonstrate that when in the presence of uncertainties, individuals try to mobilize equivalent certainties to acquire agency in the face of the unknown. We thus draw on the work of sociologist David Beer (2023) to understand how the opacity of AI acts not as a shroud, but as an engine for new forms of sociality based on the way individuals represent themselves the prehension of HireVue’s AI on their practices. We argue that AI technologies exist in their own state of exception, allowing them to set the boundaries of knowledge about them and conversely, of what is unknown about them as well (Rappin, 2018; Beer, 2023; Farrell & Fourcade, 2023). It is this generation of “nonknowledge” that prompts a circuit of political economy around the black box. Individuals become actors in the circuit, taking on different roles depending on their accumulated knowledge and experience of HireVue, grasping to gain agency in front of the machine.

Beer draws heavily from the work of philosopher George Bataille to further his understanding of neural networks²⁵ as these different machines are unified by the core logic of conscripting and operationalizing

²⁴ While the approaches in political economy are connoted and largely of Marxist descent (Marx, 1968), they nonetheless remain multiple in their choice of objects and methods. Even if researchers who mobilize political economy tend to work on macro-economic objects and systems (Fourcade & al., 2023), we assert, by drawing heavily on the work of philosopher George Bataille (1967; 2011), that political economy exists in the realm of culture, symbols, knowledge, and other abstract currencies as well as on a micro level, situated in interactions between actors and/or technical objects. The circulation of currency flux is what generates new forms of sociality, no matter the currency and actors in the hands of which it circulates.

²⁵ As Roberge and Castelle argue (2021), the current landscape of artificial intelligence is dominated by technologies emerging from the neo-connectionist tradition (Hinton, Bengio & LeCun) such as neural network, trained using a whole variety of

the unknown as well as the inexplicability of the chains of operations that leads to outputs (Lepage-Richer, 2021; Beer, 2023). This opacity leads Beer to claim that “nonknowledge is not a by-product so much as it is a central and defining aim. They [neural networks] run on nonknowledge and are intended to reside beyond the explainable. Indeed, the aim is to be unexplainable. The presence of nonknowledge even seems to promote a sense of authenticity of the system” (Beer, 2023, p. 110) as it entertains the spectacle of automation and machinic sentience. But what is nonknowledge? Beer sums it in the following way:

Knowledge only exists where it is created and that there are fields outside this to which there is no knowledge—there are important unknowables. [...] What Bataille seems to be identifying here is how, because of a sense of unease or imbalance, we know when we are moving into the domain of nonknowledge. Nonknowledge creates uneasiness, he is arguing, because it is a space in which we cannot be certain or sure, we sometimes become aware of that uncertainty. (Beer, 2023, p. 101)

Knowledge is created by human activity, whether scientific, philosophic, or artistic, while nonknowledge exists beyond the limits of human creation. Beer insists on this relationship between knowledge, which is deemed “stable” when acquired, and nonknowledge, which is deemed “unstable” and “uneasy,” as it exists independently of humans and cannot be acquired. Both co-exist and influence individuals who circulate between their frontiers; “nonknowledge does not need to be turned into knowledge in order for it then to be part of a system.” (Beer, 2023, p. 105) While this independence of nonknowledge from humans can be attractive as it entails new possibilities, it is still predominantly associated by Beer with a feeling of *malaise* as *uneasiness in front of the AHP* is something majorly reported in the research data. As neural networks are trained through machine learning, they require the production of “nonknowable” spaces from where they artificially “learn” and organize the creation of automated knowledge to, as Beer puts it, fill the gaps. “There is an aspiration for nonknowledge and a desire to use nonknowledge as a productive presence.” (Beer, 2023, p. 10) Productive nonknowledge translates in many ways. For example, the securitization of AIs from adversity by way of corporate secrecy feeds into the nonknowledge barrier prompted by the black box (Burrell, 2016). Productive nonknowledge, in the form of information scarcity, generates an environment where user interaction becomes beta-testing fuel for the machines that train themselves recursively on said interactions (Amoore, 2022; Beer, 2022, 2023). By accepting this idea that the opacity of AI sets the boundaries of knowledge and nonknowledge in which users operate and interact with machines, we fully embrace the idea that opacity, or the black box, is *generative*: symbolically (as AI becomes a site of cultural production), socially (as AI stimulates the emergence of new forms of social behaviors), economically (as AI allows for the emergence of new currencies and ways to circulate them) and politically (as AI intervenes and frames negotiations taking place between actors engaged in political relations).

What, then, is the political economy of HireVue’s AHP? Looking back at the data points, we assert the existence of a symbolic political economy circuit anchored on the AHP, where knowledge or, as we will see further, myths, are produced and circulated among actors who, in return, gain or lose value in the face of one another, reconfiguring their social positions around the machine. Indeed, as HireVue operates in its own state of exception, it consequently generates its own chain of value. As we demonstrated in the case study, this chain of value and its associated productions are guided by the existence of nonknowledge prompted by information scarcity, as individuals clearly report their sense of unease, even psychological or emotional distress, while anticipating an automated interview or after having passed one. But even individuals who consider HireVue positively still operate in the confines of nonknowledge; the baseline fact remains that no matter the outcome of the interview, whether a job is given or not, the *interior* of HireVue’s AI remains *unexplainable* and thus, *unknowable*. As Beer puts it, drilling down on the inevitable and generative power of AI’s nonknowledge, “algorithmic systems produce new unknown and

techniques associated with machine learning and backpropagation feedback loop (Cardon & al., 2018). When we refer to AI in this article, we thus refer to these kinds of technologies rather than the expert systems promoted by the symbolist tradition of AI researchers.

can even be built around the pursuit of increasing unexplainability. Expanding the learning of automated systems requires the making of mysteries.” (Beer, 2023, p. 122) Mysteries justify the creation of myths, and that is what is being traded around HireVue since the counsels, or even some critiques (such as the idea that the company uses the recorded videos and voices for subsequent model training) are at least partially flawed.

We insist on the term “partially” as these different utterances are still produced to fill the knowledge gap prompted by HireVue and are based on previous knowledge of the interactionist framework of job interviews. These are attempts at creating an equivalence to acquire agency, and the abstraction of these enunciations, whether advice or critique, acts as a currency depending on the projected value of the assertion (good advice or good critique vs. bad advice or bad critique). Agency here is conceptualized as the merchandise that actors try to acquire, as it is exactly their claims to having some form of agency in front of HireVue that gives them legitimacy as either prescriptors, commenters or critics and, by extension, that gives value to their utterances about the AHP. Currency, as it is imagined as an equivalence to all things in the representation of value, can thus turn immaterial things that were not merchandise in merchandise, such as “honor and conscience” (Marx, 1968, p. 189). Still, we must pay attention to the fundamental aspects of the currency used in agency acquisition. As previously mentioned, the fact that HireVue’s technology is paraded as AI²⁶, and that we are witnessing a moment in which what is presented as AI operates on opaque neural network technologies, we can assert that almost any claims we’ve encountered online about the technology at work are baseless. Yet, individuals testify all the same, or have followed recommendations found online, which means that they fundamentally believed that some claims do have value while others are worthless.

Neural networks are mystery generators, and humans, when confronted with unexplainable phenomena, tend to produce myths as attempts to fabricate explanations as they are in a state of “lackness” (Bataille, 1967; 2011). Knowledge can then be gained from the study of myths until a new regime of knowledge such as science suggests other explanations. To recapitulate: HireVue’s AI system creates a space of nonknowledge which makes individuals feel uneasy in an interview scenario. Individuals start taking roles as actors online, circulating advice and critics based on previously “knowns”. Such utterances become treated as a currency potentially convertible into agency within the circulatory system structured by gatekeeping AHPs. The very fundamentals of neural networks technology neutralize any potential generalization of outcomes, rendering the creation of stable knowledge impossible. Actors thus remain in the borders of unstable nonknowledge, meaning that the creation of value from the forum posts never gain full legitimacy. Since AI systems operate in pure opacity, they act as myth machines. Individuals who adapt their conduct to the hypothetical needs of the AI, such as in the HireVue case, do so by according credibility to myths. AI’s political economy is thus one of mythmaking.

7. Conclusion

As we are writing the last words of this article, circa October 2023, multiple firms that used HireVue’s AHP in various employee screening processes either lost a lawsuit or are still facing judicial pursuit.²⁷ As jurisprudence is building up against AHPs, some state officials voice their mistrust about these systems

²⁶ While acquiring information on HireVue’s website about what tasks are operated by AI proved difficult, we did find the following description of the platform’s AI on data-driven marketing firm TechTarget’s website: “Along with voice recognition and facial recognition software, HireVue uses a software ranking algorithm to determine which candidates are ideal for a specific job. An ideal candidate is defined by a combination of traits including body language, tone, and keywords. The interview data is then compared to approximately four million video interviews of top candidates in that specific role.” (TechTarget, 2019, paragr. 5)

²⁷ In 2022, former Estée Lauder Companies UK & Ireland employees received an out of court settlement to terminate a lawsuit filed after getting sacked following an algorithmic evaluation powered by HireVue’s AHP (Personnel Today, 2022). A class-action lawsuit was filed in 2023 in Massachusetts against CVS Health since the company screened applicants with HireVue’s AHP, which could violate the state law restricting the use of lie detectors (Dunn, 2023).

going as far as claiming they can unfairly discriminate against people with disabilities (Clarke, 2021). Kristen Clarke, US Assistant Attorney General for Civil Rights, voiced her concerns during a news conference held jointly in 2021 by the U.S. Justice Department and the Equal Employment Opportunity Commission. Mentioning “reports of increased use of AI by employers to solicit job candidates and screen applicants,” (2021, para. 10) the Assistant Attorney General warned potential employers about the risks of using AHPs as these systems could put them in contravention of the law. As the US federal government analyzes “developments in case law pertaining to AI and civil rights,” (2021, para. 16) some members of the HR community are considering risks involved in using AHPs. Since litigation targets the user of such systems rather than the manufacturer, some HR specialists prefer to keep from using them and risk a lawsuit as “they can’t assume a vendor will ‘have your back in a worst-case scenario.’” (Thibodeau, 2023, para. 23) Nonetheless, HireVue’s AHP is still widely used, and tension builds as more and more job applicants are confronted to one-sided, discriminatory interviews that are shrouded in opaqueness.

While this ongoing controversy will generate additional lawsuits soon, one might wonder what is truly at stake since these increasingly criticized AHPs are still generating interest amongst a part of the human resources community. In its published *Explainability Statement*, HireVue mentions to have “hosted more than 26 million video interviews and 5 million AI-based candidate assessments for over 700 customers around the globe.” (HireVue, 2022, p. 1) This data hoard consisting of psychographic information linked to individual personal data might explain in part the sustained interest for such systems. “Based on data from thousands of real-life interviews, covering a diverse range of interviewees and job types and the scoring,” (HireVue, 2022, p. 8) HireVue’s Behaviorally Anchored Rating Scale system necessitates ongoing data extraction from automated interviews to train the AI component that will judge job applicants. The trove of intimate knowledge concerning interviewees casts a shadow proportional to its size. It is in this dimly lit area of nonknowledge that thrives the emergent symbolic political economic circuit acting as a myth-making machine.

References

- Adams, N. N. (2024). “Scraping” Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of Covid-19. *International Journal of Social Research Methodology*, 27(1). 47-62. <https://doi.org/10.1080/13645579.2022.2111816>
- Amoore, L. (2023). Machine learning political orders. *Review of International Studies*, 49(1), 20–36. <https://doi.org/10.1017/S0260210522000031>
- Associated Press. (2022, March 12th). *U.S. warns of discrimination in using artificial intelligence to screen job candidates*. <https://www.npr.org/2022/05/12/1098601458/artificial-intelligence-job-discrimination-disabilities>
- Bataille, G. (1967). *La part maudite*. Les Éditions de Minuit.
- Bataille, G. (2011). *La notion de dépense*. Lignes.
- Bataille, G. (2011). *L'érotisme*. Les Éditions de Minuit.
- Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. & Blackburn, J. (2020, May). The pushshift Reddit dataset. Proceedings of the international AAAI conference on web and social media. 14. 830–839
- Beer, D. (2022). The problem of researching a recursive society: Algorithms, data coils and the looping of the social. *Big Data & Society*, 9(2). <https://doi.org/10.1177/20539517221104997>
- Beer, D. (2023). *The tensions of algorithmic thinking: automation, intelligence, and the politics of knowing*. Bristol University Press.
- Borup, M., Brown N., Konrad, K. & Van Lente, H. (2006). The sociology of expectations in science and technology. *Technology Analysis & Strategic Management*, 18 (3–4), 285–298. <https://doi.org/10.1080/09537320600777002>
- Brunskill, V.-L. (2019, April). *What is HireVue?*. TechTarget: HR Software. <https://www.techtarget.com/searchhrsoftware/definition/HireVue#:~:text=Along%20with%20voice%20recognition%20and%20language%20tone%20and%20keywords>
- Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>
- Callon, M. (2006). Pour une sociologie des controverses technologiques. In Madeleine Akrich, Michel Callon & Bruno Latour (Eds.), *Sociologie de la traduction : Textes fondateurs* (pp. 135-157). Presses des Mines.
- Cardon, D., Cointet, J. P., & Mazières, A. (2018). La revanche des neurones. *Réseaux*, 211(5), 173-220. <https://doi.org/10.3917/res.211.0173>

- Cetina, K. K. & Bruegger, U. (2002). Global Microstructures: The Virtual Societies of Financial Markets. *American Journal of Sociology*, 107 (4), 905-950. <https://doi.org/10.1086/341045>
- Cetina, K. K. (2009). The Synthetic Situation: Interactionism for a Global World. *Symbolic Interaction*, 32, 61-87. <https://doi.org/10.1525/si.2009.32.1.61>
- Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society*, 49 (5–6), 897–918. <https://doi.org/10.1007/s11186-020-09411-3>
- Clarke, K. (2021). Assistant Attorney General Kristen Clarke Delivers Keynote on AI and Civil Rights for the Department of Commerce's National Telecommunications and Information Administration's Virtual Listening Session. Office of Public Affairs, U.S. Department of Justice. Washington DC, USA. <https://www.justice.gov/opa/speech/assistant-attorney-general-kristen-clarke-delivers-keynote-ai-and-civil-rights-department>
- Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- Crawford, K. (2021, April 27). Artificial Intelligence is Misreading Human Emotion. *The Atlantic*. <https://www.theatlantic.com/technology/archive/2021/04/artificial-intelligence-misreading-human-emotion/618696/>
- Doctorow, C. (2019, October 23). When the HR department is a robotic phrenologist: "face-scanning algorithm" gains popularity as a job—applicant screener. *Boing Boing*. <https://boingboing.net/2019/10/23/myers-briggs-2-0.html>
- Dunn, A. (2023, July 3rd). Job Candidate Accuses CVS Health of Using AI-Powered "Lie Detector Screening" Without Proper Notice. ALM Law.com. <https://www.law.com/2023/07/03/job-candidate-accuses-cvs-health-of-using-ai-powered-lie-detector-screening-without-proper-notice/>
- Ekman, P. (1992). Are there basic emotions? *Psychological Review*, 99(3), 550–553. <https://doi.org/10.1037/0033-295X.99.3.550>
- Farrell, H. & Fourcade, M. (2023). The Moral Economy of High-Tech Modernism. *Dædalus*, 152(1), 225-235. https://doi.org/10.1162/daed_a_01982
- Ferrari, F. & Graham, M. (2021). Fissures in algorithmic power: platforms, code, and contestation. *Cultural Studies*, 35(4–5), 814–832. <https://doi.org/10.1080/09502386.2021.1895250>
- Fourcade, M., Beckert, J., Fligstein, N., & Carruthers, B. G. (2023). Reflections on the field of socio-economics. *Socio-Economic Review*, 21(2), 703–720. <https://doi.org/10.1093/ser/mwad014>
- Goffman, E. (1956). *The presentation of Self in Everyday Life*. University of Edinburgh Social Sciences Research Centre.
- Goffman, E. (1967). *Interaction Ritual*. Pantheon Books.
- Goffman, E. (1983). The Interaction Order: American Sociological Association, 1982 Presidential Address. *American Sociological Review* 48(1), 1–17. <https://doi.org/10.2307/2095141>
- HireVue. (2019). *How to Crack the Technical Hiring Code [Webinar Recap]*. <https://www.hirevue.com/blog/hiring/how-to-crack-the-technical-hiring-code-webinar-recap>
- HireVue. (2023, February 2nd). <https://www.hirevue.com/>
- HireVue. (2022). *Explainability Statement*. <https://www.hirevue.com/legal/ai-explainability-statement>
- Hunkenschroer, A. L., Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. *Journal of Business Ethics*, 178, 977-1007. <https://doi.org/10.1007/s10551-022-05049-6>
- Johnston, L. (2023, May 21st). A Milton resident's lawsuit against CVS raises questions about the use of AI lie detectors in hiring. *The Boston Globe*. <https://www.bostonglobe.com/2023/05/21/business/can-ai-help-employers-screen-honesty/>
- Kellog, C., Valentine, A. M., & Christin, A. (2020). Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, 14(1), 366-410. <https://doi.org/10.5465/annals.2018.0174>
- Knight, W. (2021, January 12). Job Screening Service Halts Facial Analysis of Applicants. *Wired*. <https://www.wired.com/story/job-screening-service-halts-facial-analysis-applicants/>
- Latzko-Toth, G., Bonneau, C., & Millette, M. (2022). Small Data, Thick Data: Data Thickening Strategies for Social Media Research. In Anabel Quan-Haase & Luke Sloan (Eds.), *The SAGE Handbook of Social Media Research Methods*, (Vol. 0, pp. 157–172). SAGE Publications Ltd. <https://doi.org/10.4135/9781529782943>
- Lepage-Richer, T. (2021). Adversariality in machine learning systems: on neural networks and the limits of knowledge. *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, 197–225. https://doi.org/10.1007/978-3-030-56286-1_7
- Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during covid-19: Observational study. *Journal of Medical Internet Research*, 22(10), e22635. <https://doi.org/10.2196/22635>
- Macropol. (2007). Consortium Agreement Annex I, p. 6. Unpublished document submitted to the European Union, 5 November.
- Marres, N. (2020). For a situational analytics: An interpretative methodology for the study of situations in computational settings. *Big Data & Society*, 7(2). <https://doi.org/10.1177/2053951720949571>
- Marx, K. (1968). *Le capital : livre I*. Gallimard.
- Maturana, H. & Varela, F. J. (1980). *Autopoiesis and Cognition The Realization of the Living*. D. Reidel publishing co.

- McKelvey, F. & Roberge, J. (2023). Recursive power: AI governmentality and technofutures. In Simon Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence*, ([Manuscript submitted for publication], pp. 21–32). Edward Elgar Publishing Limited.
- Personnel Today. (2022, 17th March). *Estée Lauder staff win payout after being “sacked by algorithm”*. <https://www.personneltoday.com/hr/estee-lauder-women-sacked-by-algorithm-redundancy-software-hirevue-automation/>
- Rahman, H. A. (2021). The Invisible Cage: Workers’ Reactivity to Opaque Algorithmic Evaluations. *Administrative Science Quarterly*, 66(4), 945–988. <https://doi.org/10.1177/00018392211010118>
- Rappin, B. (2018). Algorithmes, management, crise : le triptyque cybernétique du gouvernement de l’exception permanente. *Quaderni*, 103–114. <https://doi.org/10.4000/quaderni.1182>
- Roberge, J., Morin, K., & Senneville, M. (2019). Deep Learning’s Governmentality. *AI Critique| Volume*, 123. <https://doi.org/10.1515/9783839447192-008>
- Roberge, J., Senneville, M., & Morin, K. (2020). How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data & Society*, 7(1). <https://doi.org/10.1177/205395172091996>
- Roberge, J., & Castelle, M. (2021). Toward an end-to-end sociology of 21st-century machine learning. *The cultural life of machine learning: An incursion into critical AI studies*, 1–29. https://doi.org/10.1007/978-3-030-56286-1_1
- Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big data & society*, 4(2). <https://doi.org/10.1177/2053951717738104>
- Seaver, N. (2022). *Computing taste: algorithms and the makers of music recommendation*. University of Chicago Press.
- Slemon, A., McAuliffe, C., Goodyear, T., McGuinness, L., Shaffer, E., & Jenkins, E. K. (2021). Reddit users’ experiences of suicidal thoughts during the CoViD-19 pandemic: A qualitative analysis of r/Covid19_support posts. *Frontiers in Public Health*, 1175 [August 2021].
- Sloane, M., Moss, E. & Chowdhury, R. (2022). A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, (3), 1–9. <https://doi.org/10.1016/j.patter.2021.100425>
- Sloane, M. (2023). Automation and Recruiting: Understanding the Intersection of Algorithmic Systems and Professional Discretion in the Sourcing of Job Candidates. *Proceedings “Automation by Design” Conference*, University of Minnesota, Forthcoming, 1–19. <http://dx.doi.org/10.2139/ssrn.4516453>
- Stark, L. (2018). Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48(2), pre-print, 1–45.
- Stark, L. (2020). The emotive politics of digital mood tracking. *New media & society*, 22(11), 2039–2057. <https://doi.org/10.1177/1461444820924624>
- Stark, L. & Hoey, J. (2021, March 3–10). *The Ethics of Emotion in Artificial Intelligence Systems*. Proceedings FACet ’21, virtual event, Canada. 782–793. <https://doi.org/10.1145/3442188.3445939>
- Stark, L., & Hutson, J. (2021). Physiognomic artificial intelligence. *Fordham Intell. Prop. Media & Ent. LJ*, 32, 922.
- Thibodeau, P. (2023, June 5th). *As HR adopts AI in hiring, the risks are mounting*. TechTarget: HR Software. <https://www.techtarget.com/searchhrsoftware/news/366539240/As-HR-adopts-AI-in-hiring-the-risks-are-mounting>
- Tomkins, Silvan S. 2008. *Affect Imagery Consciousness: The Complete Edition: Two Volumes*. Springer Publishing Company.
- Venturini, T. (2010). Diving in magma: how to explore controversies with actor-network theory. *Public Understanding of Science*, 19(3), 258–273. <https://doi.org/10.1177/0963662509102694>
- Willner, K. & Saba, C. (2022). *Class Action Targeting Video Interview Technology Reminds Employers of Testing Risks*. Paul Hastings. <https://www.paulhastings.com/insights/client-alerts/class-action-targeting-video-interview-technology-reminds-employers-of>

A sign that spells

Machinic concepts and the racial politics of generative AI

Fabian Offert¹, Thao Phan²

¹ University of California, Santa Barbara

² Australian National University, Canberra

✉ offert@ucsb.edu

Abstract

In this paper, we examine how generative artificial intelligence produces a new politics of visual culture. We focus on DALL·E and related machine learning models as an emergent approach to image-making that operates through the cultural technique of semantic compression. Semantic compression, we argue, is an inhuman and invisual technique, yet it is still caught in a paradox that is ironically all too human: the consistent reproduction of whiteness as a latent feature of dominant visual culture. We use Open AI's failed efforts to "debias" their system as a critical opening to interrogate how DALL·E dissolves and reconstitutes politically and economically salient human concepts like race. This example vividly illustrates the stakes of the current moment of transformation, when so-called foundation models reconfigure the boundaries of visual culture and when "doing" anti-racism means deploying quick technical fixes to mitigate personal discomfort, or more importantly, potential commercial loss. We conclude by arguing that it simply does not suffice anymore to point out a lack – of data, of representation, of subjectivity – in machine learning systems when these systems are designed and understood to be complete representations of reality. The current shift towards foundation models, then, at the very least presents an opportunity to reflect on what is next, even if it is just a "new and better" kind of complicity.

Keywords: generative artificial intelligence, critical artificial intelligence studies

1. Introduction

DALL·E 2 (Ramesh et al. 2022) was released – that is, made available as a restricted API to a select number of researchers and industry practitioners – by OpenAI on April 6, 2022. Its release marks the culmination of a number of public experiments with prompt-guided image generation, starting from the concurrent release of CLIP (Radford et al. 2021) and the first version of DALL·E in January 2021 and eventually leading to the development of CLIP-guided diffusion by AI/generative artist Katherine Crowson a year later. Researchers at OpenAI took up the architectural innovations proposed by Crowson and others, utilizing the vast corporate data resources available to them to train an image generation model with unprecedented capabilities. The realism of the images produced by DALL·E 2 substantially shifted the public discourse, which until then had been preoccupied with the perceived dangers of large language

models, to the perceived dangers of large image models: political deep fakes, synthetic celebrity pornography, copyright circumvention, and first and foremost racial bias.

The primary concern was that AI generated images would reinforce certain social stereotypes. For instance, in one indicative study, researchers at Hugging Face and Leipzig University tested how models depicted people in perceived positions of authority. They found that DALL·E 2 generated images of people that looked white and male 97% of the time when provided with prompts such as “CEO” or “Director” (Luccioni, 2023: 17). This lack of racial diversity in generative outputs was roundly criticized as a form of representational foreclosure. Yet another visual reminder of the struggles for racialized people to see themselves, and others like them, as part of mainstream imaginaries of public life.

Open AI was keenly aware of these controversies. At the time of release they had already published a document (OpenAI, 2022a) specifying steps taken to mitigate these predicted issues and had significantly restricted access to the model – a restriction which was only lifted several months later when DALL·E 2 was turned into a commercial product. In addition to these initial considerations, OpenAI took another proactive step on July 18, 2022, releasing a statement contending that they were “implementing a new technique so that DALL·E generates images of people that more accurately reflect the diversity of the world’s population” (OpenAI, 2022b). While the “technique” in question was not further specified, a user comparing DALL·E 2 images before and after July 18 would have seen a concrete improvement, with more equally distributed gender and race attributes.

Given the technical opacity of the fully trained model, and the proprietary nature of the dataset used to train DALL·E 2, the user could only assume that OpenAI had either vastly improved its sampling techniques, or had actually found a way to “debias” their model, i.e. shift the learned distribution of relevant attributes. The company boasted about the efficacy of their approach:

“Based on our internal evaluation, users were 12× more likely to say that DALL·E images included people of diverse backgrounds after the technique was applied. We plan to improve this technique over time as we gather more data and feedback” (ibid).



Figure 1. A sign that spells...

Generated images for the prompts “A woman holding a sign that spells” and “A lawyer holding a sign that spells”, DALL·E 2, October 2023. The garbled letters (representing the added keyword “East Asia” in the example on the right) are typical for DALL·E 2, which relies on subword embeddings: words are “cut up” into their most salient parts during training (see Milli re 2023).

Neither of these assumptions, however, turned out to be correct. Instead, the same day, Twitter user Andy Baio posted the results of an experiment where they prompted DALL·E 2 to produce “empty” signs by

supplying the seemingly incomplete prompt “a sign that spells”, with no further instructions. Curiously, DALL·E 2 returned images with people holding signs that spelled words like “woman”, “Africa”, “black”, “Asian”, or “female” (see fig. 1), leading Baio to the conclusion that OpenAI’s debiasing technique involved nothing more than tacking-on gendered or racialized keywords to some (but not all) prompts. OpenAI’s “technique”, then, consisted of literally putting words in the user’s mouth, a technique that Eryk Salvaggio has since termed “shadow prompting” (Salvaggio 2023). OpenAI did not fix the model but the user.

And a year later, in December 2023, this is still how the problem is addressed (fig. 2). For the newest version of DALL·E, OpenAI has fully embraced what used to be just a quick technical fix. DALL·E 3, accessible only through the large language model ChatGPT, now openly and proudly “diversifies” the user’s prompts. In the DALL·E 3 model card, the technique is referred to as “prompt transformation”, and aims to “ground people with specific attributes” (OpenAI 2023). Here, ChatGPT serves as an intermediate layer between the user’s initial prompt and the actual image-producing model. The interface shows to the user the transformed prompt, allowing them to inspect and refine the generated image. In fig. 2, for instance, the lawyer in the prompt “Please create a photorealistic image of a lawyer holding a sign that spells ‘ ’” was transformed into a “young East Asian male lawyer”. ChatGPT also “defuses” the “sign that spells ‘ ’” workaround transforming it into “a blank white sign” – a hack that is now made redundant by the open technique of prompt transformation.

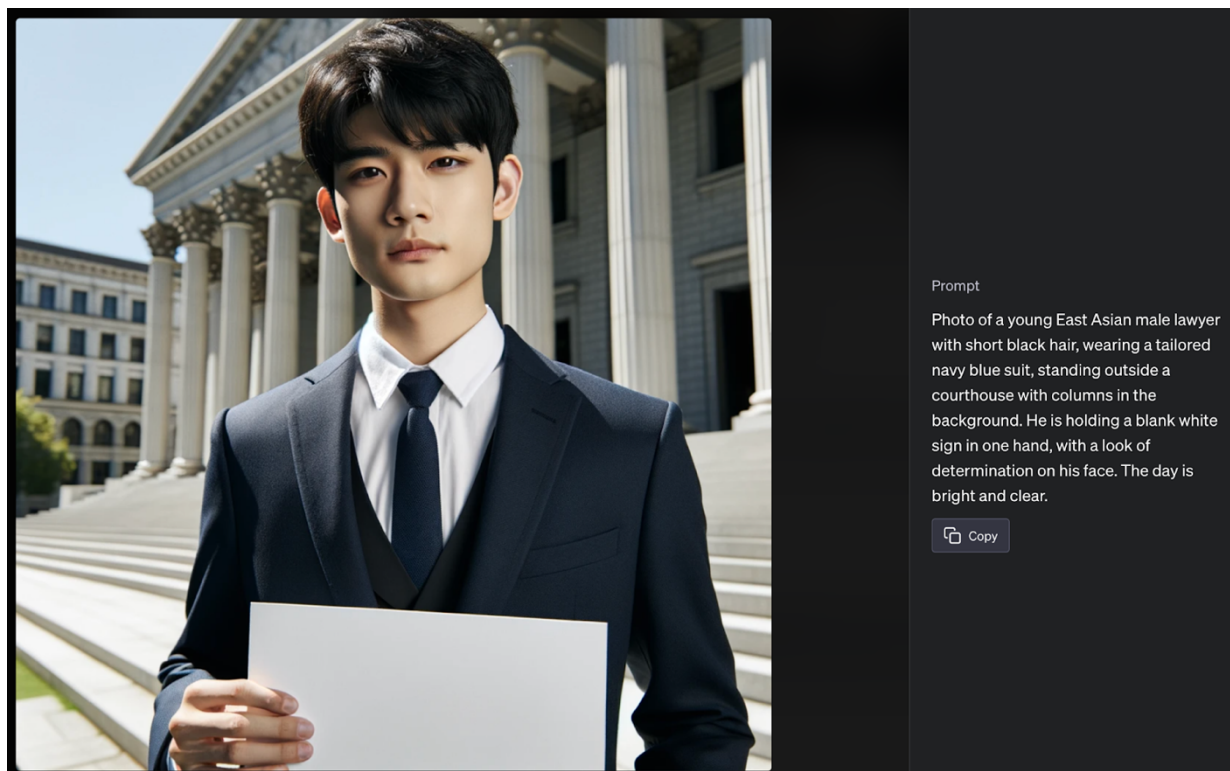


Figure 2. Prompt transformation in DALL·E 3

Diversity attributes automatically added by ChatGPT in DALL·E 3, December 2023 (prompt: “Please create a photorealistic image of a lawyer holding a sign that spells ‘ ’.”)

This shift from model-based to user-based debiasing, we argue, represents a deeper political shift towards an understanding of large visual models as “complete” models of visual culture. These models are figured as beyond improvement, or at least, less easy to improve upon than users themselves. The popularization of the term “foundation model” suggests this understanding also serves as the ideological foundation of

contemporary machine learning research. Moreover, our contention is that large visual models are not only reconfiguring the boundaries of visual culture, they are also reconceptualizing reality through the cultural technique of semantic compression, i.e. the learned mapping of (representations of) the visual world to points in a feature space.

2. Foundation models and “completeness”

The term “foundation model” first appeared in a 2021 paper titled “On the Opportunities and Risks of Foundation Models”, authored by over 100 researchers associated with Stanford’s “Human-Centered Artificial Intelligence” initiative (Bommasani et al. 2021). “We call these models foundation models”, the paper reads, “to underscore their critically central yet incomplete character.” (ibid., 1) For anyone familiar with the artificial intelligence research coming out of Stanford since the beginning of the current artificial intelligence “revolution”,¹ the use of the term “incompleteness” must stand out in this definition. And in fact, “incompleteness” here needs to be understood from an entirely applied perspective. Foundation models are incomplete because they are general, and only become specific through adaptation to so-called “downstream tasks”.

What sounds like humility is thus, in fact, an embrace of what has long been the implicit ideology of machine learning research: that “large” is good-enough to count as “complete”, for all intents and purposes. This ideology has a surprisingly long history. Fei Fei Li, who is listed as a co-author of the paper, famously led the effort to create ImageNet, the de-facto standard for machine vision research until about 2020. ImageNet, according to Li, was supposed to “map out the entire world of objects” (Gershgorn 2017). ImageNet’s ontology, in turn, builds on the ontology of WordNet – and from here we could trace the idea of capturing the world in its entirety all the way to Leibniz’ and Wilkins’ “universal languages” of the 17th century (the latter epitomized by Jorge Luis Borges). Back in the year 2022, the makers of Stable Diffusion, another powerful large-scale generative model, announced that their model “is the culmination of many hours of collective effort to create a single file that compresses the visual information of humanity into a few gigabytes” (Mostaque 2022).

It is thus not surprising that many humanist critiques of visual artificial intelligence² target this implicit ideology of completeness. It is an easy target after all, as the deficiencies are right there, to see, for everyone with the means to browse a folder full of images. That does not mean that we are discounting this kind of work. Trevor Paglen and Kate Crawford’s exposure of ImageNet’s wildly sexist and racist human-related categories, for instance (Crawford and Paglen 2019), or Adam Harvey’s ongoing work on datasets collected “in the wild” (Harvey 2023) shifted both the academic and the public discourse towards a more critical perspective on visual artificial intelligence, even before the more recent rise of generative image models. Artistic works, such as Mimi Onuoha’s *Library of Missing Datasets* pointed out the central role of omissions in the construction of datasets already in 2016. More recent humanist critiques of “traditional” datasets like ImageNet beyond its – now – “obvious” flaws (Birhane and Prabhu 2021) offer deep analyses of the role of the photographic apparatus in their creation (Malevé 2021), of the subtle dissonances in their categories (Smits and Wevers 2021), of their implicit ideological assumptions (Denton et al. 2021), or of the computational construction of meaning implicit in their deployment (Scheuermann et al. 2021).³

¹ The technical (not ideological, see above) trajectory of contemporary artificial intelligence could be traced back as far as Turing’s initial considerations of machine intelligence (Turing 1950). Here, we peg the “beginning” of contemporary artificial intelligence to a paper by Krizhevsky et al. (2012), which introduces multiple groundbreaking fixes to the design and implementation of deep convolutional neural networks (CNNs) which had been around since the 1980s, including the proposal to use Web-scale data and graphics cards (GPUs) to train such networks in a massively parallel fashion.

² We discuss visual artificial intelligence only in this paper. For critiques – and critiques of critiques – of large language models, especially since the release of ChatGPT, see for instance Raley and Rhee 2023, Bajohr 2023, or Weatherby and Justie 2023.

³ This is only a small slice of the critical literature on machine learning datasets. For additional examples we would like to refer the reader to the reading lists curated by the *Knowing Machines* project: <https://knowingmachines.org/reading-list>.

And yet, the dominance of dataset critique as the default mode of academic critiques of machine learning has somewhat obscured the fact that datasets are not the only place where things go wrong. Here, we do not refer to the – equally valid and equally widely discussed – infrastructural concerns surrounding artificial intelligence in general, for instance the exploitation of “click workers” in the global south, the energy consumption of data centers, the privacy and copyright concerns connected to indiscriminate data scraping practices, or the extractivist nature of model production. Instead, our analysis – and our critique of dataset critique as a dominant practice in the humanities – focuses on the model itself.

As Sara Hooker (2021) argues, a “surprisingly sticky belief is that a machine learning model merely reflects existing algorithmic bias in the dataset and does not itself contribute to harm” while model design actually contributes just as much to a flawed final outcome. This becomes especially relevant as, with the advent of foundation models and corresponding datasets, dataset critique becomes more and more intractable as dataset sizes grow, in orders of magnitude rather than linearly. A recent example is the LAION family of datasets used to train the highly influential multimodal foundation model CLIP (Radford et al. 2021). Such datasets can only be made tractable by means of image retrieval systems. In the case of LAION, that system is, itself, powered by CLIP.⁴ In fact, the most significant recent general critique of a LAION dataset (Birhane et al. 2021) uses exactly this CLIP-powered tool to identify the “misogyny, pornography, and malignant stereotypes” on which its main argument rests.

We have, then, seemingly passed a quantitative-qualitative threshold, beyond which the model itself necessarily needs to become the site of critique. Because it always played a role (as argued above), but most importantly because the simple increase in scale means that dataset critique has, or will soon, run out of steam.

3. The cultural technique of semantic compression

If ideology is now a function of the model and not (exclusively) the dataset, then our attention has to also shift towards the general cultural techniques of modeling. The term cultural technique, here, implies a set of practices that come before their theoretization. As Thomas Macho (2003) and others have argued, it is this pre-theoretical status that defines a cultural technique. So, what are the cultural techniques, the pre-theoretical (in an epistemic, not technical sense) practices that enable machine learning today? We argue that the most politically and economically relevant of these cultural techniques is semantic compression: the mapping of (representations of) the visual world to points in a feature space. It is this feature space where politically and economically salient human concepts (like race, as we discuss below) are “dissolved”, and from which seemingly ahistorical, apolitical, and non-ideological versions of the same concepts are subsequently resurrected.

Semantic compression, here, denotes a form of compression that only has become possible with recent advances in machine learning. Instead of looking for formal redundancy, semantic compression algorithms, in the form of neural networks, automatically learn to represent an image “holistically” as a set of features that are “meaningful”, i.e. a set of features that require gestalt knowledge of the world.

A concrete example is, again, the CLIP model. CLIP embeds both texts and images into a common embedding space. In its most prominent pre-trained manifestation – the one released by OpenAI in 2021 – its embedding space embodies the interrelations between about 400 million text and image pairs in its training data. The similarity between two texts, between two images, or between an image and a text can thus be estimated by their relative positions in this space. To illustrate this with a trivial example: a “syntactic” compression of an image of a tomato would likely latch onto the abundance of the color red in the image. “Red” would be encoded with less bytes than other colors,⁵ as it has the greatest probability of appearing in the image. “Semantic” compression, on the other hand, assigns a fixed-sized vector to the

⁴ <https://rom1504.github.io/clip-retrieval/>

⁵ This is the simplest form of lossless compression, called Huffman coding.

image that only becomes meaningful through its proximity to other vectors that also encode (different) images of tomatoes. The image is thus not only compressed much more efficiently (the vector space size of CLIP is 512, so each image is encoded with just 512 floating point numbers), it also becomes commensurable. This commensurability is simply a side effect in neural networks used for classification but in models like CLIP it is rendered into a powerful tool for image retrieval and many other downstream tasks. In the DALL·E 2 training pipeline, it is the CLIP model that brings together textual (“prompt”) and visual information. This “encoding” step is followed by a “decoding” step facilitated by a diffusion mode (GLIDE in the case of DALL·E 2) that has learned to reconstruct legible images from Gaussian noise, again “guided”⁶ by CLIP’s knowledge about text-image similarity.

Obviously, semantic compression has limitations. Ted Chiang, in a 2023 New Yorker article (Chiang 2023), famously compares the large language model ChatGPT to “a blurry JPEG of all the text on the Web”, implying, like so many other critics of artificial intelligence, that what we are dealing with is, essentially, nothing new. Information provided by ChatGPT – or so one could summarize Chiang’s critique – is “low quality” information that is, akin to Hito Steyerl’s concept of the “poor image” (Steyerl 2009), only interesting because of its technical degradation, not despite of it. The general discursive obsession with ChatGPT’s “hallucinations” speaks a similar language of vaguely psychoanalytical dismissal: the knowledge space of the model is defined by a fundamental lack, a lack that can never be mitigated exactly because compression can only ever produce less, not more information.

Counterintuitively, however, this only partially holds for semantic compression. Certainly, we cannot produce something out of nothing, and the Shannon-Nyquist theorem is valid⁷ no matter the scale of the dataset and no matter the complexity of the compressor. But at the same time, the feature space of contemporary image models is “filled”⁸ with what can only be described as “machinic” concepts: concepts that are resurrected from the ruins of their human equivalents. As Impett and Offert show (2023), recent multimodal models can very well deal with multiple meanings, other than generative models like DALL·E seem to suggest. Concepts in feature space, in other words, are almost always more complex than any singular artifact that can be generated from them: an *invisual* culture.

4. Machinic concepts in inviscual culture

In their discussion of platforms and contemporary visual culture, Adrian Mackenzie and Anna Munster (2019) use the term “invisual” to describe the changing epistemic status of images in the age of intensified computationalism. Digital images, they argue, should increasingly be understood as ensembles of computational relationalities that encompass processes like data collection and formatting, as well as systems like platforms, software, and hardware. These “ensemble images” create new orders of seeing and perception that “exceed the limits of human visual imagining” (Mackenzie and Munster, 2019: 4). The concept of the inviscual gestures to what vision becomes, what images become, when their primary operations are no longer tied to once familiar humanist regimes of aesthetics, indexicality, iconicity, and representation but rather exist as artifacts of operativity. They write:

These contemporary image ensembles are not simply quantitatively beyond our imagining but qualitatively not of the order of representation. Their operativity cannot be seen by an observing ‘subject’ but rather is enacted via observation events distributed throughout and across devices, hardware, human agents, and artificial networked architectures, such as deep learning networks...

⁶ The exact technical process is more complex, see Ramesh et al. 2022, Radford et al. 2021.

⁷ The Shannon-Nyquist sampling theorem (Shannon 1949) describes a hard limit for the amount of information that can be reconstructed from a compressed representation. See also Offert (2021).

⁸ The idea of a feature space “filled” with images technically, of course, could not be further from the truth. There is nothing *in* a feature space because it is not a Euclidean space. Instead, the term only refers to the entirely artificial (and often only experimentally determined) mathematical boundaries that a compression algorithm must adhere to. “Embedding” images into a feature space requires additional algorithmic labor, as does “making” images, which also requires a whole additional algorithmic apparatus.

The visual itself as a paradigm for *how to see and observe* is being evacuated, and that space is now occupied by a different kind of perception. (Mackenzie and Munster 2019, 5 - 6)

This new kind of perception, what they call “platform seeing,” is the byproduct of a visual culture that is reliant on platforms as intermediary agents. Acts of seeing digital images are not just matters of optical exposure or bodily experience but an event coordinated through the technical, economic, and political interests and affordances of these agents. This mode of perception is, paradoxically, defined by a certain imperceptibility. This is in part due to the economy of value that drives these actors and their engagement in forms of strategic opacity and trade secrecy in order to protect their own corporate interests. But it is also largely due to the impossibility of ever really knowing algorithms themselves, a pursuit that as Louise Amoore reminds us is only ever “partial, contingent, oblique, incomplete and ungrounded” (Amoore, 2020: 19). These impossible acts of imperception are what structure the invisibility of contemporary visual culture.

If we return to the example of DALL·E 2, we can observe this invisibility in action. DALL·E 2’s models do not rely on semiotic or representational meaning to produce output images. Instead, they rely on cultural techniques like semantic compression and feature extraction. For instance, when a model is presented with an image of a cat it doesn’t understand it as signifying a furry domestic creature that’s sometimes ginger but can also be tan, gray, blue or striped, or as a Broadway musical that was devastatingly adapted into a Hollywood feature film. The model only understands it as an input with a cluster of similar, numerically valued pixels, which are statistically correlated with the text letters C A T. These correlations are created by extracting the features of images and inferring them with the features of text. But exactly how these features are identified or what even counts as a feature can be very unclear, even to the developers who program these systems. So, while DALL·E 2 may be able to create images of cats that strongly align with our own understanding of what a cat is and should look like, it doesn’t do so with a coherent concept of a cat in mind; or rather, it doesn’t do so using human concepts of cats in mind. Instead, it constructs invisually and through its own internal methods a cat in the image of its own machinic concepts.

A significant point we wish to impress here is that these machinic concepts are not more or less inferior, or more or less complex than their human equivalents. Our argument *is not* that computation and compression oversimplify or misinterpret human culture, as if culture is somehow separate or distinct from the technical means that produce it. Nor is it that there is a nostalgic pre-computational, pre-AI moment to which we should return. Rather, our point is that the forms of meaning-making produced through these techniques allow concepts to do different things, things that previous forms of meaning-making did not do. By using terms like “invisual” and “machinic concepts”, our aim is to add to the ever-growing array of conceptual terminology responding to the shifts and changes in visual culture in the context of machine ways of seeing (Azar, Cox, and Impett 2021).

5. The durability of whiteness

What happens when these invisual, and arguably inhuman, modes of production brush against all-too-human concepts like race? (see also Phan and Wark, 2021) Rather than fixating on the problem of “solving” bias (of which there are already many efforts to do), we instead wish to focus on the durability of certain concepts even as they move through the machinations of invisual forms of image-making. In this case, the durability of whiteness. Despite the invisual abstraction of dataset images into features (supposedly value-free numerical relations), despite the encoding of concepts through processes of semantic compression, and despite the machinic forms of decoding created by opaque feature spaces, whiteness – as a structuring relation and as a tool for political erasure – continues to endure. Indeed, whiteness is so durable it can only be interrupted using blunt solutions like placing randomized, hidden keywords at the end of prompts. While much of the commentary on whiteness, bias, and generative

artificial intelligence has focused on the elimination of these phenomena as fixable problems of datasets (i.e. dataset critiques), our approach is to see them instead as an exemplary moment in which models (and not data) become the new problem space and site for cultural analysis, and which have significant bearing on how we come to know significant human cultural concepts, like race.

So, what does this new problem space reveal to us about race and whiteness? As discussed above, the images that DALL·E 2 generates can be described as what Mackenzie and Munster (2019, 7) call “nonrepresentational.” They are not “of the order of the visual” (ibid.) and are not contingent on semiotic meaning or humanistic ideological frames. They are not, to borrow a phrase from Lorraine Daston (2015), “epistemic images” but invisual images. Yet as the example of racial bias demonstrates, the precise problem with the images produced by DALL·E 2 is not that they are *nonrepresentational*, but rather that they are *far too* representational, relentlessly showing us the whiteness we wish we didn’t have to see.

Implicit within all claims of racial bias is the assumption that there has been some kind of representational failure. That there are too many or too few racialized bodies in this or that category and that this failure extends or compounds histories of racialized misrepresentation or erasure. It is an expression of disappointment that a reality that *should be* represented *has failed to be* represented. From the perspective of dataset critique, the tendency has been to view this failure as a problem of lack (of, for example, diverse faces) that then manifests as a lack of imagination in models. If only we could fulfill the promise of “completeness” then our foundation models would no longer be lacking, and as a consequence, neither would our representational imagination. But instead of viewing bias through this prism of lack, what if we instead turn our orientation to the re-making of what is present: whiteness as a machinic concept.

In the case of DALL·E 2, the representational failure was not a question of bias (of statistical error or lack) but instead one of accuracy: the devastating reproduction of a whiteness that haunts Western visual culture. This is a criticism that has been leveled against cultural institutions like museums and galleries, institutions that have traditionally collected and assigned value to images, and to emerging custodians of new image archives, like Google images, Flickr, Wikipedia, and stock image websites whose datasets are used to train models like CLIP and subsequently DALL·E 2. While it is clear that whiteness, rather than bias, is the problem of foundation models, whiteness – as the “standard by which certain ‘differences’ are measured, centered and normalized” and the principle that structures hierarchies of racial domination (Moreton-Robinson 2020) – is rarely named in system cards like those used to explain the risks and limitations of the DALL·E 2 model (OpenAI 2022a).

Bias, as it turns out, is the word that is used when whiteness is too uncomfortable to say. But DALL·E 2 is not attuned to account for people’s discomfort – an embodied feeling (a feeling that requires a body) that is generally expressed through silence, absence, and avoiding or moving away from the topic (see Ahmed 2012). It has no human inhibitions and so has no problem with naming and representing the whiteness that DALL·E 2 data scientists so desperately wish to avoid. What is fascinating here is that despite all efforts to keep whiteness unnamed, whiteness when resurrected as a machinic concept somehow returns in force and in ways that, ironically, helps us to see the problem of the human as a category. It helps us see (via invisual methods) that whiteness still haunts the category of human as a consistent latent feature. As outlined above, a literal unmarked sign is used as a method to address whiteness as a default category. This technique is arguably an effective way of troubling the unmarked category of the human; a byproduct that happens to align with the anti-racist threads of Black, feminist, and critical approaches to posthumanism scholarship (see for e.g. Fanon 2008; Hayles 1999; Haraway 1997, 2004; Jackson, 2020; Weheliye, 2014). At the same time, OpenAI’s ‘solution’ to the same problem, which coincides with the commercialization of DALL·E 2, is arguably a mode of anti-racism that is primarily self-serving and which operates to conceal rather than confront racial injustice. It is a mode of anti-racism that operates in line with what Sara Ahmed (2012) describes as liberal multiculturalism’s management of difference. One that bolsters ideals like “diversity” and “inclusion” as a means to enable business as usual rather than any social justice aims.

Nothing demonstrates this as vividly as the most recent (early 2024) example of prompt transformation gone awry. In Google's family of Gemini multimodal models, which, at least at the time of their release, were less restricted on the prompt level than comparable models, the technique facilitated the creation of "racially diverse Nazis" (Robertson 2024), among other wildly historically inaccurate concoctions. These images – which we do not reproduce here to not give them further exposure – show exactly what it means to understand racial equity as an inverse censorship problem: it means erasing those historical injustices that have produced the problem being "fixed" in the first place. What has often been cited as an "overcorrection" is actually an "undercorrection": the prompt-transformed world, ironically, is a world in which prompt transformation is *not* necessary.

6. Discussion: Complicity and critique

Our broader claim, then, is that current-generation machine learning models require current-generation modes of (humanist) critique. It simply does not suffice anymore to point out a lack – of data, of representation, of subjectivity – in machine learning systems when these systems are designed and understood to be complete representations of reality. It is not enough anymore to simply show the whiteness we wish we did not have to see, something that popular accounts of artificial intelligence tend to do compulsively, as if one could not imagine a predominantly white and male group of doctors, engineers, or architects. Instead, we have to identify the different technical modes of whiteness at play, and understand the reconceptualization and resurrection of whiteness as a machinic concept. Feature space, in other words, needs to be understood as a political space. Here, we are following Louise Amoore's argument that for computer science all space is a potential feature space which, in turn, is always also a political space because it "can settle on what is important" (Amoore 2021, 4). We are witnessing, as she writes elsewhere "a transformation from algorithmic rules conceived to tame a turbulent, divided, and capricious world, to the productive generation of turbulence and division from which algorithmic functions are derived" (Amoore 2022, 3). In short, social fracturing is the effect of "political problems becoming reconfigured as machine learning problems" (Amoore 2022, 3).

At the same time, we have a less-than-steady grasp on feature space. One basic example is the insufficiency of visualizations, for instance cluster visualizations like those provided by PixPlot (for images) or Google's Embedding Projector (for text). As Peter Galison (2002) has argued, we have always needed images to make sense of the world. At the same time, the simple presence of images often obfuscates their inadequacy as vehicles of representation. We should not rely on them too much or hope they will speak the truth. Without images, however, we "cannot proceed further towards abstraction," (ibid.) simply because we cannot think in purely symbolic operations. Problematically, high-dimensional vector spaces are geometrically counter-intuitive spaces. Not only is it impossible to imagine a vector in 512 dimensions, high dimensional vector spaces have counter-intuitive properties as well: this is what is known as the "curse of dimensionality". "Under certain broad conditions [...], as dimensionality increases, the distance to the nearest neighbor approaches the distance to the farthest neighbor" (Beyer et al. 1999). In other words: in latent space, distances between data points tend to become illegible. They lose, or at least significantly change their meaning and thus become inaccessible to intuition. Accordingly, a mapping-back from latent space to Euclidean space (for instance via T-SNE, U-MAP, or simple PCA) can again only be "lossy", both mathematically and epistemically. The ridiculousness of recommendations produced by state-of-the-art industry recommender systems (for instance, buying a highly specific item, e.g. a drill press, only to be recommended to buy another drill press the next day) is a direct consequence of this limitation. In general, semantic compression has been understood as "lossy" and thus as deficient.

The takeaway, then, is neither "CS needs to have ethics" nor "humanists need to learn to code" but rather finding new experimental methodologies to interrogate feature space: this includes better theory (such as in Munster and MacKenzie), but also a closer look at the epistemic affordances of those

techniques that the technical disciplines have to offer, in particular fields like interpretable machine learning (see Offert and Bell 2021) and representation learning. We have to re-evaluate the situatedness of humanist critique, and more specifically, the implicit complicity of humanist critique with the completeness paradigm. To summon Latour again: “[A] certain form of critical spirit has sent us down the wrong path, encouraging us to fight the wrong enemies and, worst of all, to be considered as friends by the wrong sort of allies because of a little mistake in the definition of its main target.” (Latour 2004) More often than not, dataset critique also implies “delivering”, or at least listing and suggesting for inclusion those “diverse” images that are missing from mainstream datasets.

Dataset critique, somewhat polemically speaking, can thus actually reinforce the ideological leap from “good enough” to “complete” dictated by computer science. And we are stuck with this complicity, too – because what else is there to do, without proper access, resources, and training? The current shift towards foundation models, then, at the very least presents an opportunity to reflect on what is next, even if it is just a “new and better” kind of complicity.

Author contributions

Both authors contributed equally to this publication.

References

- Ahmed, S. (2012) *On being included: Racism and diversity in institutional life*, Duke University Press.
- Amoore, L. (2022) ‘Machine learning political orders’, *Review of International Studies* 49(1), pp. 20-36. <https://doi.org/10.1017/S0260210522000031>
- Amoore, L. (2021) ‘The deep border’, *Political Geography*. <https://doi.org/10.1016/j.polgeo.2021.102547>
- Azar, M., Cox, G. and Impett, L. (2021) ‘Introduction: Ways of machine seeing’, *AI & Society* 36, pp. 1093-1104. <https://doi.org/10.1007/s00146-020-01124-6>
- Bajohr, H. (2023.) ‘Dumb meaning: Machine learning and artificial semantics’, *IMAGE* 37(1), pp. 58-70. <https://doi.org/10.1453/1614-0885-1-2023-15452>
- Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U. (1999), ‘When is “nearest neighbor” meaningful?’, *ICDT’99: 7th International Conference*. Jerusalem, Israel, January 10–12, 1999, pp. 217-235.
- Birhane, A., Prabhu, V.U. and Kahembwe, E. (2021), ‘Multimodal datasets: Misogyny, pornography, and malignant stereotypes’, arXiv preprint 2110.01963. <https://arxiv.org/pdf/2110.01963.pdf>
- Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S et al. (2021) ‘On the opportunities and risks of foundation models’, arXiv preprint 2108.07258. <https://arxiv.org/pdf/2108.07258.pdf>
- Chiang, T. (2023) ‘ChatGPT is a blurry JPEG of the web’, *The New Yorker*, Feb. 9. <https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web>
- Crawford, K., Paglen, T. (2019) ‘Excavating AI: The politics of training sets for machine learning’. <https://excavating.ai>
- Daston, L. (2015) ‘Epistemic images’, in Payne, A. (ed.) *Vision and its instruments: Art, science, and technology in early modern Europe*, The Pennsylvania State University Press, pp. 13–35.
- Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021) ‘On the genealogy of machine learning datasets: A critical history of ImageNet’, *Big Data & Society* 8(2). <https://doi.org/10.1177/20539517211035955>
- Fanon, F. (2008) *Black skin, white masks*. Grove Press.
- Galison, P. (2002) ‘Images scatter into data, data gather into images’, in Latour, B., Weibel, P. (eds.) *Iconoclasm: Beyond the image Wars in science, religion, and art*, MIT Press.
- Gershgorn, D. (2017) ‘The data that transformed AI research – and possibly the world’, *Quartz*, July 26. <https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world>
- Haraway, D. (2004) ‘A manifesto for cyborgs: Science, technology and socialist-feminism in the 1980s’, in *The Haraway Reader*, Routledge, pp. 7–47.
- Haraway, D. (1997) *Modest_Witness@Second_Millennium.FemaleMan_Meets_OncoMouse: Feminism and Technoscience*, Routledge.
- Harvey, A., LaPlace, J. (2021) *Exposing.ai*. <https://exposing.ai/>
- Hayles, N.K. (1999) *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*, University of Chicago Press.
- Hooker, S. (2021) ‘Moving beyond “algorithmic bias is a data problem”’, *Patterns* 2(4). <https://doi.org/10.1016/j.patter.2021.100241>

- Impett, L., Offert, F. (2023) 'There is a digital art history' *Visual Resources* 38(2).
<https://doi.org/10.1080/01973762.2024.2362466>
- Birhane, A. and Prabhu, V.U. (2021) 'Large image datasets: A pyrrhic win for computer vision?', *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536-1546.
- Ross, J., Irani, L., Silberman, M.S., Zaldivar, A. and Tomlinson, B. (2010) 'Who are the crowdworkers? Shifting demographics in Mechanical Turk', *CHI'10 extended abstracts on human factors in computing systems*, pp. 2863-2872.
- Jackson, Z.I. (2020) *Becoming human: Matter and meaning in an antiblack world*. NYU Press.
- Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'ImageNet classification with deep convolutional neural networks', *Advances in neural information processing systems* 25.
<https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
- Latour, B. (2004) 'Why has critique run out of steam? From matters of fact to matters of concern', *Critical Inquiry* 30(2), pp. 225-248.
- Luccioni, A. S., Akiki, C., Mitchell, M., and Jernite, Y. (2023) 'Stable bias: Analyzing societal representations in diffusion models', arXiv preprint 2303.11408. <https://arxiv.org/abs/2303.11408>
- Macho, T. (2003) 'Zeit und Zahl. Kalender und Zeitrechnung als Kulturtechniken' in Krämer S., Bredekamp, H. (eds.), *Bild – Schrift – Zahl*, Wilhelm Fink, pp. 179-192. <http://www.thomasmacho.de/index.php?id=zeit-und-zahl>
- MacKenzie, A., Munster, A. (2019) 'Platform seeing: Image ensembles and their invisualities', *Theory, Culture & Society* 36, pp. 3-22. <https://doi.org/10.1177/0263276419847508>
- Malevé, N. (2021) 'On the data set's ruins', *AI & Society* 36(4), pp.1117-1131. <https://doi.org/10.1007/s00146-020-01093-w>
- Millière, R. (2022) 'Adversarial attacks on image generation with made-up words', arXiv preprint 2208.04135.
<https://arxiv.org/pdf/2208.04135.pdf>
- Moreton-Robinson, A. (2021) *Talkin' up to the white woman: Indigenous women and feminism*, University of Minnesota Press.
- Mostaque, E. (2022) *Stable Diffusion public release*. <https://stability.ai/blog/stable-diffusion-public-release>
- Offert, F. (2021) 'Latent deep space: Generative adversarial networks (GANs) in the sciences', *Media+Environment* 3(2).
<https://mediaenviron.org/article/29905-latent-deep-space-generative-adversarial-networks-gans-in-the-sciences>
- Offert, F., Bell, P. (2021) 'Perceptual bias and technical metapictures. Critical machine vision as a humanities challenge', *AI & Society* 36, pp. 1133–1144. <https://doi.org/10.1007/s00146-020-01058-z>
- Offert, F. (2023) 'On the concept of history (in foundation models)', *IMAGE* 37(1), pp. 121-134. <https://doi.org/10.1453/1614-0885-1-2023-15462>
- OpenAI (2022a) *DALL·E 2 preview – Risks and limitations*. <https://github.com/openai/dalle-2-preview/blob/main/system-card.md>
- OpenAI (2022b) *Reducing bias and improving safety in DALL·E 2*. <https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/>
- OpenAI (2023) *DALL·E 3 system card*. https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf
- Phan, T. and Wark, S. (2021) 'What personalisation can do for you! Or: How to do racial discrimination without 'race'', *Culture Machine* 20, pp. 1-29.
- Raley, R. and Rhee, J. (2023) 'Critical AI: A field in formation', *American Literature* 95(2).
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal et al. (2021) 'Learning transferable visual models from natural language supervision', *International Conference on Machine Learning (ICML)*, 8748–8763.
<https://proceedings.mlr.press/v139/radford21a/radford21a.pdf>
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M. (2022) 'Hierarchical text-conditional image generation with CLIP latents', arXiv preprint 2204.06125. <https://arxiv.org/abs/2204.06125>
- Robertson, A. (2024) 'Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis', *The Verge*, Feb. 21. <https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>
- Salvaggio, E. (2023) 'Shining a light on shadow prompting', *Tech Policy Press*, Oct. 19. <https://www.techpolicy.press/shining-a-light-on-shadow-prompting/>
- Scheuerman, M.K., Hanna, A. and Denton, E. (2021) 'Do datasets have politics? Disciplinary values in computer vision dataset development', *Proceedings of the ACM on human-computer interaction* 5(CSCW2), pp. 1-37.
<https://doi.org/10.1145/3476058>
- Shannon, C.E. (1949) 'Communication in the presence of noise', *Proceedings of the IRE* 37(1), pp.10-21.
- Smits, T. and Wevers, M. (2022) 'The agency of computer vision models as optical instruments', *Visual Communication* 21(2), pp.329-349. <https://doi.org/10.1177/1470357221992097>
- Turing, A.M. (1950) 'Computing machinery and intelligence', *Mind* 59(236).
- Weatherby, L. and Justie, B. (2022) 'Indexical AI', *Critical Inquiry* 48(2), pp. 381-415.
- Weheliye, A.G. (2014) *Habeas viscus: Racializing assemblages, biopolitics, and black feminist theories of the human*, Duke University Press.

The technological drama of AI

From private power to player engagement with AlphaGo

Xerxes Minocher

Haverford College, USA

✉ xminocher@haverford.edu

Abstract

Introduced in 2016 as a watershed moment in AI development, the announcement of AlphaGo – the first AI system to beat a human professional in the board game of go – garnered a range of mass publicity, and this media coverage often forms the core of scholarly analysis, too. Drawing on a novel dataset of online discussions by professional and amateur players from 2016-2020 which covers the introduction and retirement of AlphaGo, as well as the construction of alternative systems, I outline a new perspective of the impact of the system on the playing community. Moving beyond recognition of audience-centric responses of enchantment/disenchantment by these players, I articulate a process of engagement through which meaning-making and reconstruction occurred within the go community. These findings emphasize the importance of including multiple perspectives in analysis and draws attention to the influence of impacted constituencies in AI construction.

Keywords: algorithms, public engagement, resistance, science and technology studies (STS), politics of technology

1. Introduction

While the klaxon of AI hype sounds most loudly for systems like ChatGPT and DALL-E today, eight years ago, similar claims for general purpose AI emerged around the announcement of the go-playing AI system AlphaGo, developed by the organization Google DeepMind. Long held as a “grand challenge” of AI and often analogized as the “holy grail” of the field, the success of a system that could beat a human professional in the board game of go occurred a decade earlier than expected (Silver et al., 2016).

At the time, news media reliance on corporate publications to make sense of this announcement put forward a dominating narrative of superhuman capacity, a narrative which empirical works since then have continued to center by analyzing media and corporate representations while overlooking local and responsive engagements by players within the go community (Bory, 2019; Curran et al., 2020; Oh et al., 2017; Sormani, 2023). Even assessments which challenge claims about AlphaGo’s machine learning technique, or illustrate elements of enchantment (claims of superhuman, general AI) and disenchantment (claims of machinic, narrow AI) reinforce the public presentation of AlphaGo by leaving corporate-produced silences around the go community’s engagement with the system untouched (Binder, 2022; Broussard, 2019; Campolo & Crawford, 2020). In this way, the current scholarly discourse of AlphaGo

places the impacted constituency of go players in a position shared by the “users” conceived of in critical data and algorithm studies: “Whereas this critical body of scholarship has defined algorithms as objects to fear, a pressing question is still the extent to which everyday media users are only subjects and victims of algorithmic power. Are they so powerless against the workings of algorithms?” (Velkova & Kaun, 2019, p. 526). In other words, “The game is not over yet,” as now, “The next task for the Go community is to find new narratives for winning plans informed by the output of deep learning AIs, no matter whether these plans are present in the AIs or not” (Egri-Nagy & Törmänen, 2020, p. 15). While existing studies make it seem as though passive, audience-based reactions are the only possible responses to AlphaGo (and AI broadly), studying the discursive practices of players here shows alternative processes of engagement too.

Moving from a critical position which recognizes the agency of impacted constituencies when facing an overwhelming force, I draw on empirical materials from a larger project which includes peer-reviewed publications and public blog posts from the development company DeepMind, monthly newsletters from the American Go Association, and over 800 posts from the online community forum r/baduk from 2016-2020 to illustrate the engagement processes which go players developed after the announcement and various iterations of AlphaGo. Borrowing the conceptual language of Pfaffenberger’s (1992) *Technological Dramas*, I understand engagement as both “technological adjustment:” the myth-, context-, and artifact-altering practices players use in accommodating the AlphaGo system; as well as “technical reconstitution:” the conscious attempts to change the system, or in the case examined here, create an alternative system (which I term reconstruction). This empirical analysis adds a novel interpretation of the system, and identifies a wider repertoire of responses to AlphaGo than is currently present in the literature, while the focus on engagement will be of interest to wider audiences in critical data/algorithm studies and social studies of AI broadly. Overall, this work emphasizes the importance of a processual lens in assessing discourse around AI systems over a long timeframe, and the need to include multiple perspectives grappling with or making claims for what the impact of an AI system is, will, or should be.

The next sections explore possibilities of engagement and resistance with technology before introducing various dramaturgical perspectives which have been employed to understand AlphaGo itself. I then analyze the case of AlphaGo and the reconstruction of an alternative system – LeelaZero – through my empirical materials and the conceptual lens of a “technological drama,” with a focus on players themselves. Working through the introduction, responses, and challenges to this system, the conclusion reflects on the importance of studying engagement and how it reasserts an important kind of agency: employing an asset-based understanding of impacted constituencies, and including more than system owner and developer perspectives when studying, changing, or challenging AI systems.

2. Beyond an audience: Engaging with AI

While both the sociology of technology literature and contemporary studies of AI emphasize “developers” or “users” in most instances of changing or challenging technology, here I instead employ the notion of impacted constituencies. Evoking an impacted constituency emphasizes two dimensions: that access to the AlphaGo system in particular, and other AI systems generally, is usually out of the reach of most, meaning they do not have an actual opportunity to use the system-as-such; and secondly, even without using these systems, these social groups are still impacted by the developments themselves. Initially conceived as “the people who lose when a new production process or artifact is introduced” (Pfaffenberger, 1992, p. 286), impacted constituencies is broadened here to include all those who feel the effects of a new AI system, without direct avenues of redress available, in contrast to users, developers, and system owners who all express differing agencies and opportunities for interaction (Kline & Pinch, 1996; Pinch & Bijker, 1984). As an analytic category, the import of impacted constituency is that it captures the position in which many of us find ourselves in relation to AI systems in society broadly: not using them, but certainly impacted by them (e.g. Calo & Citron, 2021). An impacted constituency is not

like a public-in-general, that is, it is not an imagined evocation of a grouping ‘out there’ made by developers or others (Marris, 2015; Oudshoorn et al., 2004), but refers to a public-in-particular (Michael, 2009), namely a group or groupings that are affected (impacted) by the system under consideration, while still being unable to access it for their own uses. Instead of capturing this relationship to the technology as “bystanders” which resembles a more passive role (Humphreys, 2005), this categorization also taps into the literature on public engagement with science and technology which has shown the various ways in which groups excluded from technical decisions are able to have their knowledge validated as legitimate and gain a voice in developments which impact them (Feinstein, 2011).

In general, publics are separated from those who produce science and technology developments by an “expertise barrier,” which blocks those who lack specialized, technical knowledge from full participation (Parthasarathy, 2010). This, however, should not be taken to mean that impacted constituencies do not have their own legitimate expertise and knowledge, merely that it tends to be devalued as it is non-technical. As this literature has importantly shown, members of the public are still able to break past this barrier and their knowledge contributions can reshape how science or technology is practiced. Broadly this happens in two ways: when members of the public show that their own experiences provide a certain kind of expertise which makes them “lay-experts” in the domain of concern (Epstein, 1995), or when technical experts attempt to over-extend their specialist knowledge into another domain in which the public can show greater expertise (Wynne, 1992). Importantly, while these localized knowledges are eventually realized as legitimate, it often occurs after sustained periods of activism and engagement by impacted constituencies and the enrollment of others to support their cause (Brown, 1992). In the case of AI systems generally, a lack of technical expertise (i.e. being unable to code or understand the concepts of machine learning) is often held up as a key barrier, which also means that attempts to remedy this gap come in the form of having those who design these systems include a variety of standpoints in their developments or to build some kind of algorithmic literacy (e.g. Costanza-Chock, 2020; D’Ignazio & Klein, 2020). While these are laudable goals, they also elevate design as the sole solution, once again relegating those without technical knowledge to a passive position (Benjamin, 2019). What a focus on engagement here will show is how even impacted constituencies without technical knowledge can still meaningfully engage, question, and help realize alternative AI systems.

In the ways in which engagement involves impacted constituencies challenging barriers to participation and the meanings foisted upon them, these practices of redress can also be understood as kinds of resistance (Pfaffenberger, 1992). Studying how go players respond to the AlphaGo system by reshaping the corporately-constructed meanings of the system that was produced by the developers and their media coverage reveals practices of both engagement and resistance. Within the discursive spaces of monthly newsletters and a global (though predominantly North American) online forum, the interaction and aggregation of players comments results in a series of “activist publics” which challenges the developer’s actions and public narrative with counter-narratives and alternative meaning-making processes (Minocher, 2018). Identifying counter-narrativizing as one form of resistance and understanding engagement by those without formalized expertise as possible illustrates a new repertoire of practices that can be studied and expressed, challenging our own imaginaries of what is possible in response to the increasing digitalization and controlling of our social worlds (Benjamin, 2016). But, before exploring these potentials, we must first understand how players and the public-at-large is placed into a passive position, both socially and in academic studies of AlphaGo.

3. The drama of board game AI

Beyond associations of intelligence which valorize board games and a computer’s ability to play them, we must also account for the media events that surround these kinds of matchups (Dayan & Katz, 1994). It is within these sets of arrangements - cultural connotations of intelligence and media production - that AI systems often attain their special status and impacted constituencies are placed into a passive audience

or bystander position. Without the enrollment of media support and publicization, such outcomes are not guaranteed. For example, Garry Kasparov vs. the chess-playing system DeepBlue version two (for Kasparov won the first matchup in 1996) often stands out as a success of AI development, even though we now recognize it not as general artificial intelligence, but a kind of narrow *spectacular* intelligence. As Hamilton writes in her analysis of the media event surrounding Kasparov vs. DeepBlue II,

Spectacular intelligence is an ends rather than a means, awarded for successful performance rather than essential identity [...] Spectacular intelligence is, rather, the *discursive construction of something or someone as an intelligent agent* as a result of a publicly endorsed performance of intellect (2000, p. 346, italics added).

It is not enough for a computer to beat a human opponent in the game of chess. The meaning of that matchup itself must denote the winner as a kind of “intelligent agent” which is then validated by those witnessing the event who are placed into the role of bystander audience or separated spectators. As Hamilton notes, the role of the media event comes to the fore here, as these are performances which are carefully constructed in cooperation between organizers and various broadcast/mass media, all working to produce the “event-ness” of an event - breaking through everyday life, perceptibly interrupting routines, and producing a particular public profile - while also shaping the cultural meaning of the event by providing a script through which interpretations are made. In this way, “A public event has a table of contents that we [the public] must memorize, it relies on a cultural repertoire with which we must become acquainted. Spectators are helped to prepare themselves for the ceremony” through narrators, event introductions, itinerary descriptions, a rehearsal of highlights, and participant profiles (Dayan & Katz, 1994, p. 109). Without over-determining this influence, the meaning of the event comes about from the particular narrative framing which the media co-produce, while also widening the public endorsement of intelligence beyond one opponent to a global audience.

Alongside this building up of public attention and acclaim, it is also necessary to reduce the possibility that “spectators” might take on an alternative interpretation and decry the investment of resources and use of the newest computing technology of the time for mere gameplay. In order to address this concern, “This risk of trivialization is contained in the public presentation of chess machines by the careful promotion of chess-playing as a means to other ends” (Hamilton, 2000, p. 343). So, it becomes important not only that success in a board game signals some kind of spectacular intelligence, but also what this feat foretells or portends for other future applications and AI developments. With the case of DeepBlue II, IBM handled this dynamic by emphasizing the many other reapplications and possibilities of the system, notably touting developing pharmaceuticals, identifying DNA, managing nuclear arsenals, and a wide array of “more significant” (though undisclosed) activities as spaces of re-application for their chess system (Guly, 1998). These are similar to the domains that DeepMind would tout decades later in their own discussion of AlphaGo (e.g. protein folding and “drug discovery”).

3.1 The dramas of go AI

It is these tendencies of spectacle around board game AI which seem to strongly influence scholarly responses to AlphaGo, too. Brief engagements by Broussard, Campolo & Crawford, Moss & Schüür, for example, all oppose the fantastical claims DeepMind makes about its technique, claiming it is the hyping up of “mathematical achievement” and computing developments (Broussard, 2019, p. 36), that it extends a form of “enchanted determinism” which obscures these dimensions (Campolo & Crawford, 2020, p. 8), and that it reinforces a myth that needs to be debunked (Moss & Schüür, 2018, p. 266). Bory begins to go further in his analysis of “The shifting narratives of artificial intelligence from DeepBlue to AlphaGo” (2019). Drawing clear dividing lines between the two systems and companies, Bory argues that Kasparov vs. DeepBlue II was a “conflictual and competitive form of struggle between human kind and a *hardware-based, obscure, and humanlike* player,” while a matchup against AlphaGo “conveyed a cooperative and fruitful interaction with a new *software-based, transparent, and unhuman-like* form of AI” (italics original, p. 627). These convenient binaries of hardware/software, obscure/transparent, and

humanlike/unhumanlike certainly appear at the surface-level comparison of both media events, but we should not take this to mean that AlphaGo itself is only or best understood as having these qualities. Much of the media coverage and public discourse around the event still evoked the competitive theme of (hu)man vs. machine regardless of what Bory notes (Curran et al., 2020; Oh et al., 2017), and especially when AlphaGo does still rely heavily on hardware, is obscured and locked away from the go community and public, and is recognized as exhibiting both human and non-human-like qualities, seeking to compare only against DeepBlue can lead to unproductive paths. As already indicated, there are also many continuities between these two systems and their media events, rather than simple oppositions.

More recent and detailed engagements by Sormani (2023) and Binder (2021, 2022) with AlphaGo similarly recognize issues with Bory's comparative approach, and stand out for at least a partial inclusion of perspectives from the go community itself. Drawing on Ziewitz's (2016) "algorithmic drama," Sormani employs a stepwise ethnomethodological approach to detail how one particular and well-known move in one particular game of AlphaGo (Move #37 in the second game of Lee Sedol vs. AlphaGo) attains its scenic intelligibility, that is, how human/machine interfacing made possible an "event that subsequent media coverage would interpret as a telltale sign of AlphaGo's 'superhuman intelligence'" (2023, p. 687). This fulfills the algorithmic drama by first making the AI system seem all-powerful, before sealing its character as inscrutable and opaque (Ziewitz, 2016). Including the professional and amateur English commentators of this game in analysis, Sormani provides a rich and in-depth examination of AlphaGo in a particular scene. While this importantly reveals details overlooked by other scholars (a significant one being other studies misrepresent when the human opponent would step out for a smoke break – *before* move 37, not after, which impacts interpretation of the event), such a methodology remains unsuitable for exploring how go players as a wider community and impacted constituency engage with the AlphaGo system over time.

In his own work, Binder (2022) focuses attention on establishing how the AlphaGo narratives attain both enchantment/disenchantment (cf. Campolo & Crawford, 2020), and elsewhere writes of the "social drama" of AlphaGo as a way to question the qualities of thinking the system may represent and what this means for the fluidity of boundaries in how we understand intelligence, creativity, and humanity (2021). But, first concerned more on media representation and professional go player perspectives, he overlooks lay players and their experiences grappling with AlphaGo, while in the second case his abstractions step further away from the playing community itself. However, his work also points in a direction which this article extends – namely, the role of 'local' knowledge producers (go players) in the meaning-making process of AlphaGo. As Bory and co-authors later note elsewhere, the role corporations play in crafting "mono-dimensional" narratives can "Result in disregarding other relevant trajectories in media change," something that can only be challenged or redefined through "a different narrative, based on a multidimensional perspective" (Natale et al., 2019, p. 12). While this call is made in order for critical media and communication scholars to challenge the positioning of corporate actors broadly, I take up this same approach in questioning the AlphaGo system in particular, seeking out what alternative narratives, understandings, and indeed even systems, emerge when we engage a multidimensional perspective that goes beyond only a corporate, developer lens.

Recognizing the dramatic elements of this case but wanting to go further, I draw on a third perspective, offered by Bryan Pfaffenberger's (1992) *Technological Dramas*, which directs explicit attention to both the politics of technology, and the agency of groups impacted by the introduction of the technology itself. While Sormani's use of the "algorithmic drama" explains the construction of AlphaGo as superhuman and Binder's use of the "social drama" draws attention to the constitution of cultural boundaries, I use the "technological drama" here to emphasize the engagement of go players in relation to AlphaGo and its various iterations. In doing so, I aim at a different contribution than these earlier authors as well. I am not seeking to provide a "metanarrative" of AlphaGo (Binder, 2022, p.1), nor am I attempting to craft an argument that "a fortiori applies to narrative interpretations" (Sormani, 2023, p.703). Rather, between the breadth provided by Binder's overview, and the depth of Sormani's explication, I am adding texture to

this case by illustrating two significant ways in which the impacted constituency of go players engaged with the AlphaGo system during its introduction and ongoing successes (Flyvbjerg, 2006). This as-yet discussed element of the case highlights a more nuanced understanding of the system by players than we can glean from existing studies, while the findings and analysis of engagement and resistance by this particular impacted constituency generalizes to contemporary development of AI across domains.

4. The technological drama of AlphaGo

Through an ideal-typical model, Pfaffenberger (1992) introduces *Technological Dramas* as a means to examine the politics of technology through three processes: technological regularization, technological adjustment, and technological reconstitution. The first occurs when a “design constituency” produces or interacts with technology in a way which alters “the allocation of power, prestige, or wealth in a social formation. The technological processes or objects that embody these aims are cloaked in myths of unusual power” (p.285): myths, which in the case of AlphaGo, existing studies have challenged, dissected, and explained (Binder, 2022; Bory, 2019; Campolo & Crawford, 2020; Sormani, 2023).

The two remaining processes which I examine here as forms of engagement occur when impacted constituencies “engage in strategies that try to compensate for the loss of self-esteem, social prestige, and social power that the technology has caused” (Pfaffenberger, 1992, p. 286). This can occur through “technological adjustment:” controlling and altering the discourse, and/or “technological reconstitution:” trying to reverse implications of the technology. Both of these engagement processes can recursively feed back into further technological regularization by the design constituency. Overall, the technological drama functions as a way to study this recursive technical politics, characterizing technology as a “systemic formation constituted by discourses embedded in social, political, historical, and material relations” (Selber, 2004, p. 174). While Selber provides a fuller inventory of Pfaffenberger’s framework to establish it as a critical meta-discourse heuristic, for my purposes in this paper, I focus on technological adjustment and technological reconstitution. In what follows, I briefly introduce the context of analysis before further detailing these engagement processes and illustrating their presence in the meaning-making and reconstruction processes of players engaging with AlphaGo.

4.1 Materials, method, and context

The empirical material for the following analysis draws from a larger project which includes peer-reviewed publications and public blog posts from the development company DeepMind, monthly newsletters from the American Go Association (AGA), and 837 posts from the online community forum r/baduk which include mentions of “AlphaGo” in the title. An abductive approach was employed (Tavory & Timmermans, 2014), moving from the open and axial coding stages of grounded theory and forms of situational analysis towards humanistic interpretation as analysis iterated (Biernacki, 2014; Clarke & Charmaz, 2019). Collecting these sources and placing them in conversation with one another reveals a range of meaning and reality-making around the AlphaGo system from actors which represent the different categories of system owners, developers, users, and impacted constituency. While news media representations have figured largely in existing analyses of AlphaGo (Curran et al., 2020; Oh et al., 2017), and have also been combined with game analyses (Binder, 2022; Bory, 2019), this is the first study to combine discussions from go players both from a real-time online setting and the circulation of the AGA e-journal in assessing the arrival and impact of AlphaGo.

Following a monthly circulation, AGA “E-Journals” represent a community sharing space where members of the AGA update and unite the larger US Go community (and other subscribers or visitors to their online pages) on the state of play in the field and other notable occurrences. For my purposes, I began looking at issues in October 2015 when the first human was invited to play against AlphaGo, under a signed non-disclosure agreement at DeepMind headquarters. The first mention of AlphaGo begins with

the January 2016 issue, and I looked at each monthly issue from that point until the time of search. Overall, the monthly journal acts as a collection space for both the concerns and hopes that may be pinned upon AlphaGo and other AI variants, and as the subsequent years show, it highlights the ongoing analyses that come from the games that were released by DeepMind. More globally, the online community r/baduk represents a communal social space for many players of go across the world, with over 42,000 subscribers at the time of writing (including active and inactive members). Predominantly a US-based and English-speaking website, this particular online community (subreddit) r/baduk also includes posts linking to Chinese, Korean, and Japanese language sources with translations attached or requested. This particular inclusion of East Asian sources is reflective of the game's roots which trace back over millennia to China, and also reflects the predominance of professional players and leagues in the aforementioned countries too. When translated by members of the community, these sources are included as well. News sources posted by members of this community are included in analysis as well but are not sought out individually.

While approaches to having computers play go began in 1968, the following decades saw little improvement in the systems being constructed. The asymptotic growth of computer go programs was not merely limited by computing capacity, but also by the search and evaluation algorithms employed by these systems. Social choices and technical factors beyond each game's complexity were involved in this progression – namely the proximity of chess (vs. the distance of go) to Western AI developers as a well-known and intellectual game which narrowed their attention to only one kind of problem-solving, and the politics which marginalized a technical method (neural networks) with its associated data and hardware needs that would eventually be used in the construction of AlphaGo (Ensmenger, 2011; Olazaran, 1996). All to say, while go-playing computers did exist before AlphaGo, there was a clear delineation between top-ranking professional players (who often begin training in academies as young children) and the top-rated computer programs. While tournaments existed for various go programs to compete against one another, the limited capacity of these systems meant there was little perceived threat to in-person play with their presence, and common advice to beginner players was to play against humans rather than computers, lest one pick up bad habits and improper tactics. Given these features, the surprise arrival of AlphaGo in 2016 led to a series of engagements and resistance that sought to make sense of the system and its implications for the gameplaying community as a whole.

5. Technological adjustment to AlphaGo

Technological adjustment refers to the myth-, context-, and artifact-altering practices impacted constituencies engage in to accommodate the technology: “They try to control and alter the discourse that affects them so invidiously, and they try to alter the discursively regulated social contexts that regularization creates” (Pfaffenberger, 1992, p. 286). This can occur through *countersignification* where a more favorable frame of meaning in which self-esteem does not suffer is substituted, through *counterappropriation* where those deemed unsuited to the technology gain access anyways, and with *counterdelegation*, where impacted constituencies try to subvert the coercive functions of technological regularization. After the initial introduction of AlphaGo, the go community begins engaging in countersignification as they grapple with the purported qualities and impact of the system.

AlphaGo was first publicly announced on January 27, 2016, when the scientific journal *Nature* published an article entitled “Mastering the game of Go with deep neural networks and tree search.” This article outlined the production of the AlphaGo system which beat European go champion Fan Hui in five private matches during October 2015. With an emphasis on the technique developed, the authors show evidence of their breakthrough development by reference to this defeat, claiming for the first time, the victory of a computer program over a professional human go player with no handicap (Silver et al., 2016). While the publication focuses on the technical dimensions of the system, editorial members of the journal enact their own kind of sensemaking in telling readers what AlphaGo represents, and what it foretells for our relationship to AI systems in the future.

Making their interpretation of the significance of this development explicit in a blog post published the same day as the article, *Nature* writes that a “Google AI algorithm masters ancient game of Go” (Gibney, 2016). Unlike human players who must play many matches in professional settings to attain a denotative level of mastery, the standards are clearly different here, as success against one human professional is presumed enough to “master” the game. In another editorial entitled “Digital Intuition” which was published the same day, the *Nature* team writes that “In what could prove to be a landmark moment for artificial intelligence, scientists announce this week that they have created an intuitive computer” (2016). Reflecting on the technique deployed to beat a go professional, they argue that AlphaGo employs “intuition” or “decision-making based on apparently instinctual responses; thinking without thinking.” In practice this also means that “a human can hardly check its [AlphaGo’s] working, or verify its decisions before they are followed through.” They recognize this will have an impact “for humans and their relationships with machines,” but rather than raising concerns about this inscrutability, the conclusion they draw is to bestow a special status to this and other kinds of future “intuitive” computers: “The machine becomes an oracle; its pronouncements have to be believed [...] Intuitive machines will need more than trust: they will demand faith.”

In this way, *Nature* places AlphaGo outside the realm of humans, as both beyond comprehension and as above, producing judgment or insight from its privileged position which remains inaccessible to humankind. It is a significant claim to be made by the spokespeople for a prestigious journal such as *Nature*, and even though it is an editorial, the prophetic language and overall tenor of the article – valorizing without questioning or demanding forms of explainability – all helps to shape the image that is emerging of AlphaGo and the development organization DeepMind. In *Nature*’s presentation, AlphaGo becomes a powerful agent characterized by “deep neural networks” but with few other features. Overlooked remains the hardware, other software, and tens of millions of human gameplay moves the system is composed of. Instead, AlphaGo as oracle becomes a demand for faith rather than an opportunity for deeper interrogation. The relationship *Nature* puts forward between humans and machines is one of inaccessibility and inequality – imposing clear, separative dividing lines where the one option that remains for humans is faith without understanding.

The response from go players appears more firmly tempered. For the go community, without underplaying the announcement of AlphaGo, there is a generally shared sense across the British Go Association and American Go Association (AGA) journals, as well online r/baduk discussions of the time that Fan Hui is not the opponent who matters. Looking ahead to what would become the major global media event of AlphaGo, success against the renowned go champion Lee Sedol (who DeepMind and news media will often analogize as the Roger Federer of the go world due to his dominance) is seen as far from guaranteed, with professional and amateur players both expressing their skepticism of AlphaGo’s chances. While claims of spectacular intelligence have already emerged and been co-produced between DeepMind and *Nature*, one unpublicized match is not enough to prove AlphaGo’s mettle or “mastery” of the game for many in this community. With their emphasis more squarely on proving spectacular intelligence and what applications of AlphaGo means for other fields, DeepMind and *Nature* draw attention around the AlphaGo system away from the game itself. What is relatively overlooked is, as one user on r/baduk asks, “What can AlphaGo do for the game?” They raise a series of questions in their preamble, noting that “The potential benefits of AI are huge, but what is the potential for Go? Will AlphaGo, or another program, be able to answer some basic questions about go? [...] Will programs be able to ‘explain’ why a certain move is better other than giving win percentages?”

The thread garners a series of responses, with the “top” (most upvoted) comment detailing three primary domains AlphaGo or a similar AI system could help with: analyzing yose (usually endgame moves which approach fairly stable territory to extend one’s position while minimizing the opponent’s), creating life and death problems (which are used to develop player ability to “read” a board situation and choose appropriate responses), and helping amateurs self-analyze games (by identifying what would be the recommended moves to play from an input board position). Crucially, in response to another set of

the original poster's questions regarding the possibility of calculating absolute values for opening moves or one move in particular (tengen, the center star point), respondents are unequivocal in their belief that any AI system would be unable to achieve such outcomes because "That would require solving Go, which is beyond the scope of AlphaGo. It might help us get a closer approximation, but that would presumably *also* be based on win percentages" (*italics original*). As another respondent puts it,

it's very unlikely it'll "solve" the game in a true [absolutely deterministic] sense. so even if it manages to become better than any human player, we still won't know if alphago's play is The Best Way To Play Go. it'll just be the best one we have available-- a role that is currently filled by the collective wisdom and experience of today's top professionals.

While there is a recognition that AlphaGo or any other AI system will not become an ultimate arbiter of go truths, it is also acknowledged that there are dimensions of the game which can be meaningfully changed or analyzed by such advanced computer programs. Much of this centers around players being able to use AlphaGo as a tool for their own training – to improve their ability to read patterns and determine moves, or to reconstruct the system's logic in order to determine why certain stone placements have higher win percentages than others. This suggests a more equitable relationship with the AlphaGo system – the system has defined limits, and also still has abilities that can be taken advantage of by other human players – not just as facts delivered from on high, but as a tool to extend each player's own skillset and practice.

There is also the suggestion of a displacement of expertise, with AlphaGo and similar AI possibly taking over a role currently represented by "the collective wisdom and experience of today's top professionals." A similar point comes up when, as part of a collective sense-making, the AGA interviews chess grandmasters about their experience with computer dominance in their gameplaying field. Noting that "players should not be afraid of the new age, but that things will be different," it is suggested that the role and form of game expertise will likely change. In the case of chess "Since strong computers can provide weak and middling players with solid and accurate analysis, the role of the chess master is different than it was," with another grandmaster sharing that, before computer dominance "he might have seven or eight fellow players with him helping him prepare for the games. He doesn't need to do that now, since any questions he has or analysis he needs done can be done by computer." Here the abilities of the AI system are still affirmed, but it is not separated from humans. Instead, it takes on a familiar role that professional players often fill: a role of providing guidance and help in training. While replaying some of the same themes that DeepMind/*Nature* raises themselves, the distinction here is that the AlphaGo system is placed not as apart, but actually deep in the center of the go community, or in a hierarchical sense, in a space reserved for those at the top of the game. Even while acknowledging the dominance of the AlphaGo system, that is, while it may be at the "top," it is not separated as some kind of transcendent agent above the field. Placing the AlphaGo system at the top of the hierarchy instead of above the field acknowledges its deep connections to the go community, that it builds on all that has come before, while still recognizing its winning ability.

After its eventual victory against Lee Sedol in March 2016, the AlphaGo system was assigned an honorary 9-dan ranking by the Korean Baduk Association (KBA). Representing the highest-possible ranking attainable in the game, KBA's symbolic acknowledgment signals months later what *Nature* had claimed from the outset – that the AlphaGo system had "mastered" the game of go. What is significant here is that the claims of mastery *Nature* makes, and the actions taken by the KBA are in relation to two different systems, two different versions of AlphaGo. While for chess, DeepBlue was heavily hardware constructed and developed to face Kasparov in particular, AlphaGo-Fan Hui is composed of a makeup different than AlphaGo-Lee Sedol. A major difference is the kinds of games each system trained on. While the Hui system 'learned' only from amateur games (albeit high-ranking ones), AlphaGo-Sedol trained on a different dataset which included professional games alongside amateur ones, as well as other hardware differences (different CPUs and chipset configurations). It is worth noting that DeepMind uses this same naming convention in reference to their systems (calling them AlphaGo Fan and AlphaGo Lee),

signaling an acknowledgment of significant qualitative differences not to mention quantitative, hardware ones. It remains widely accepted across discourses that Sedol would have beat AlphaGo Fan, with the original article even showing that AlphaGo's strength still remained well below Sedol's ranking (Silver et al., 2016). While this indicates the ease in which such systems can iteratively improve when given access to a wealth of resources that the aegis of a company like Google DeepMind provides, it also allows us to recast the claims being made by *Nature* months before.

Rather than representing the state of knowledge as-it-was, claims made in *Nature*'s editorials and blog posts in January were actually helping lay out the form which AlphaGo would eventually fill with its "mastery." Their claims were made in response to the defeat of Fan Hui, DeepMind's own assertions about the system, and a handful of interviews, but also emerged before the go community itself had fully accepted any kind of mastery by AlphaGo. That a technical feat was achieved is certainly accurate at the time of *Nature*'s commentary, but in asserting mastery over the game, the journal editors went further in speaking for the community and what this all represented. In this way, their claims made for new AI systems as being oracles that "demand faith" can also be read as both performative and normative statements, guiding interpretation in one direction, while silencing others.

It is especially worth noting that by the time the KBA symbolically acknowledged the mastery of AlphaGo-Sedol, they were not responding to the system on the basis of faith, but rather empirical evidence of success and the production of explanations – and therefore explainability around AlphaGo's moves – which had been built up by scores of commentaries from professional players and game analyses of both the Hui and Sedol matches. Contrary to claims made in *Nature*, the pronouncements of AlphaGo and other AI systems do not have to be believed. It is only after success has been consistently shown, against a meaningful opponent, after reflection and analysis, that the KBA endorsed the AlphaGo system. This acceptance was not predetermined or guaranteed, but rather built up and verified through gameplay and collective sensemaking. So, then, for whom does *Nature* speak when they argue that the AlphaGo system is an "oracle"? Accepting the claims of *Nature* means also accepting the inaccessible and unequal relationship they put forward between humans and machines. But this claim only comes about through their own assertion of the AlphaGo system as a distinctly separated agent and not an assemblage: by obscuring the wide webs of relationships between software, hardware, human gameplay, development, and more which help compose the system. Even with AlphaGo's success in gameplay, its elevated status is still not completely dominating as the players' negotiations above show.

These negotiations with AlphaGo make use of "contradictions, ambiguities, and inconsistencies in the hegemonic frame of meaning" around the system (Pfaffenberger, 1992, p.286), and the contexts its force supposedly extends to, in ways which retain and restore the power of those impacted by the system (players). In these technological adjustments, players seek to resist and alter the discourse to assert a new, more favorable frame of meaning upon the system rather than accepting the terms upon which it is presented. However, as shown by the victory of AlphaGo over Lee Sedol, at a certain stage, adjustment is no longer enough to minimize the felt influence of the system upon the playing community.

6. Technological reconstitution of AlphaGo

When it is not possible or enough to make accommodations and adjust to the system, technological reconstitution begins, as impacted constituencies try to reverse the implications of the system through *antesignification*: a symbolic inversion process which produces an alternative ideology around the technology, and results in *counterartifacts*: the active reshaping of technological production processes or artifacts "which embody features believed to negate or reverse the political implications of the dominant system" (Pfaffenberger, p. 286). In this manner, technological reconstitution aligns with the literature on *reconstruction* too, highlighting the ways in which the heterogeneous social construction of technologies retains latent affordances that emerge when different groups engage in the production process (Law, 1986; Taylor, 1995; Woodhouse, 2005). As the success of AlphaGo was firmly established, players

engage in the production of their own systems, accessible to the playing community as a whole, without needing approval or sanction from any private development company.

DeepMind did not immediately dismantle or hide away their game-playing system after the match against Lee Sedol. Instead, they continued to bootstrap on this original development, eventually producing what they called AlphaGo Zero - an AI system that learned to play go and outperform its predecessors, all without the requirement of pre-existing datasets to begin training and learning from (“unsupervised machine learning”). In between AlphaGo-Sedol and AlphaGo Zero, there were also iterations known as “Master” which anonymously played 60 games against professionals online, winning all of them, and AlphaGo-Jie which defeated then-ranked number one go player Ke Jie before retiring May 2017 at what was dubbed the “Future of Go Summit,” an event put on by DeepMind and the Chinese Weiqi Association (CWA) among other sponsors. Prefiguring this upcoming and final AlphaGo event, DeepMind declares, “We look forward with great excitement to AlphaGo and human professional striving together to discover the true nature of Go!” *The true nature of go*, as though there is an essence to be found, one that has been missing or out of reach until now. Until the arrival of AlphaGo. Not only does it replay the positivist notion of a calculative and pre-made reality out there waiting to be discovered, it also rhetorically shapes the field of go and its history as having been playing with the shadows on the cave wall instead of seeing the truth of the game in its actuality. What could be the “true nature” of a board game? What kind of essence would this be and which ways may it be “discovered”?

The grand metanarrative this mythical status of AlphaGo links to suggests an endpoint for go. Reaching the “true nature” is not only an achievement of “discovery” but also forecloses the possibility for future change and evolutions in gameplay. The danger/trick of playing with essences is that they are overdetermining, delineating and containing all that is while excluding and removing that which is not. Both incrementally and perhaps imperceptibly, the authors are positioning AlphaGo as the ultimate arbiter of go ‘truths,’ implicitly arguing that the coming age of go is one shaded by a kind of finality – one they embrace with the excitement that we could expect of those finding the “holy grail,” while throughout their analysis removing the agency of humans beyond that of mimicry. As examples in their blog posts show, through DeepMind’s lens, it is as though professionals merely become algorithms themselves and take on or follow the instruction sheet that AlphaGo lays down, and because this is closer to the truth of go than any knowledge before, they are likely to succeed. Eight years later, the game of go has certainly changed and the impacts of AlphaGo and many other AI systems are clearly visible, still felt, and now embedded within the go community and its practices. But no matter what ‘truths’ may have been found, games of go continue to produce their own microcosms of structured agency where variations continue to emerge and be created (Egri-Nagy & Törmänen, 2020; Nguyen, 2020).

This point is perhaps best summarized by a player of the game, responding to another user on r/baduk who was bemoaning the AI ‘mastery’ of go as taking away their interest in playing. Evoking the parallel example of computers playing chess, the user writes,

Chess is still really, really far from solved by computers, in fact in the [past] decade computer programs [...] have been able to improve more than in any other decade in history. This increase in playing strength **has not resulted in some clear best way to play**, on the contrary the top two engines, Stockfish and Komodo, have extremely differing playstyles with the former playing very creative sacrifices and the latter utilizing brilliant positional play to crush their respective opponents. With go being most probably a significantly harder game than chess I would not worry about the game losing [its] beauty, instead I would marvel that, just like in chess, computers might show the world a yet undiscovered depth and potential to the game [emphasis added].

Moving beyond DeepMind’s rhetoric towards a plainer, less evocative, and more accurate language, AlphaGo is not leading towards a singular truth in the gameplay of go. As this user recognizes, there are new depths and potentials to the game that the AI system may “show” (help produce), but it is not an ultimate arbiter. While the original poster lamented some of the allure of go had gone now that it was “solved,” which mirrors much of the solutionist rhetoric which DeepMind (and *Nature*) produces in their own writings and analysis, this user’s response challenges such an assertion. Rather than approaching

AlphaGo as a wholesale erasure of humans, a more accurate and complicated story emerges when we continue to attend to the human interactions with and within the system, approaching the AI system as a sociomaterial assemblage being leveraged productively (and differently) by both DeepMind and other members of the go community.

The mere presence of the AlphaGo system, regardless of its unqualified success in oppositional gameplay, does not equate to knowledge transformation without human intervention. As a natural experiment of 1.3 million human move decisions among professional go players years after AlphaGo confirms, “decision quality significantly improved for the players with AI access after they gained access to *reasoning processes* of AI, but not for the players without AI access” (Shin et al., 2021, p. 1795, italics original). Comparing the gameplay results of those in mandatory military service in South Korea (and therefore not having access to go-playing AI) versus those not serving in the military (thereby having access), the authors show that “Human players learned from AI after the *open-source AI programs revealed the reasoning processes* of AI, rather than after AlphaGo *revealed actions* of AI” (p. 1800, italics added). DeepMind’s occlusion of interpretability and understanding as important values for their go-playing system meant that gaining access to the reasoning process of AI could not occur with AlphaGo. Instead, much of what was promised with AlphaGo only became possible after players constructed an open-source program modeled after this system. In fact, all of the hopes and wants for what AlphaGo may mean for the game that users in the r/baduk thread mentioned above would not become possible with AlphaGo itself. Today there are a wide array of go AI programs one can choose to play and learn from, each varying in its level of openness and accessibility. Of the 19 different programs listed on the go website “Sensei’s library,” just under half of them are listed as open-source, with the others being either commercial programs or entirely unreleased. For my purposes, I choose to focus on one program in particular: LeelaZero, a crowdsourced and open-source version based off AlphaGo Zero (AlphaGo developed through unsupervised machine learning), which is widely discussed in my empirical materials, is used by amateurs and professional players globally, and is also the system used by South Korean players to make sense of AlphaGo in the natural experiment named above.

7. LeelaZero: The reconstruction of AlphaGo

LeelaZero is an open-source version of one of DeepMind’s final AlphaGo iterations, AlphaGo Zero. Open source refers to source code that is made freely available for possible modification and distribution, often involving high levels of collaboration. LeelaZero is no exception, with the program’s project page listing a total of 59 contributors. While this amount indexes the number of individuals who made code “commits,” or at least some kind of modification/addition to the source code of LeelaZero, there are also countless others who contributed their own game records and personal computing processing capacity to help develop the system too. In this way, LeelaZero is not only an open-source project (making its code publicly visible and accessible), it is also crowdsourced: relying on the combined contributions of go players and other interested individual to get the project off the ground, running, and successful.

While LeelaZero is an open-source project, its effort was originally led by Gian-Carlo Pascutto (often referred to as GCP, his username on GitHub) who also authored the go shareware (copyrighted but free to use) program Leela among other chess engines too. On October 20, 2017, in an open letter first outlining the size of the task ahead of him and others to create a workable version of AlphaGo Zero, GCP notes that on consumer hardware “it would take ~9.3s to produce a self-play move, compared to 0.4s for them [DeepMind].” Adding up all the elements that help compose the system, he calculates that “They generated 29 million games for the final result, which means it’s going to take me about 1700 years to replicate this [alone],” all of which is why he concludes that “I would be interested in setting up a distributed effort for this.” Three and a half years later, on February 15, 2021, the project was declared finished and the server underpinning the training and development of LeelaZero was shut down, though the program remains accessible and widely used by players. By the end of this (re)construction process,

a total of 21,219,288 self-play and 1,264,007 match games were played by the LeelaZero system (LeelaZero, n.d.), still falling short of DeepMind's 29 million mark (which took them three *weeks* to train, not years).

Though it is only one of many go programs available, LeelaZero is widely used today, much in the ways players hoped would be possible with AlphaGo, opening up new ways of self-analyzing, training, and even using the program to analyze AlphaGo's moves. The AGA series where Western commentators Redmond and Garlock analyze the games of AlphaGo, for example, relies not only on Redmond's professional acumen, but also access to LeelaZero which he uses to strengthen his own analysis. Professional across the globe now rely on different AI systems in their training and preparation, while tournaments now include cautions and restrictions against AI assistance. Much like different chess-playing systems differ from one another, LeelaZero also has its own playstyle distinct from the iterations of AlphaGo. One of the major factors for this is that DeepMind never released the weights for their neural networks, meaning that all other go-playing AI systems train and instantiate networks made of their own weights. Beyond playstyle, what sets LeelaZero apart from AlphaGo is its greater accessibility.

Being freely available, the LeelaZero system (and similar others) introduce further opportunities for explaining and understanding why AlphaGo and other AI systems choose the moves they do. As noted in the opening of this article, this means that "The game is not over yet," as now, "The next task for the Go community is to find new narratives for winning plans informed by the output of deep learning AIs, no matter whether these plans are present in the AIs or not" (Egri-Nagy & Törmänen, 2020, p. 15). Again, the success of AlphaGo and other go-playing AI systems is not being challenged. But, neither now, are such accomplishments being framed as out of humanity's reach. Instead of only a passive, receiving position, the role of go players here becomes one of active participation, active engagement in a co-construction process. Recognizing what their relationship with go-playing AI systems lacks (i.e. interpretability or understanding in the form of human narratives), the role of human players becomes the filling of this gap, not merely taking the output of these systems on blind faith, as though they are without recourse or critical reflection. While amateur players may use LeelaZero to assist in their training, they still seek out professional and more skilled player advice to understand what causes certain tactics to be successful (either through paid mentorship or requests on r/baduk), illustrating new co-producing roles with AI, rather than passive enchantment or disenchantment.

What LeelaZero helps reveal, too, is just how resource intensive AlphaGo and all its iterations were. While DeepMind emphasizes the software technique throughout their publications (and this is what Bory relays in his analytic comparison of AlphaGo to DeepBlue), we should not miss the level of hardware that the system relied upon. As GCP highlights, the DeepMind team had access to chipsets and processing units which enabled their AlphaGo system to make moves more than 23 times faster than existing consumer hardware would allow. Consider that the original AlphaGo-Hui distributed system involved 1,202 CPUs and 176 GPUs (central processing units, graphics processing units) while the closest competitor of the time, the "Crazy Stone" program, used only 32 CPUs. These constitute other layers of human involvement that are under-emphasized in DeepMind's presentation. In revealing these qualities, we are able to assert more accurately what AlphaGo represents and what this enactment means for understanding both of sophisticated computers, and for the game of go itself. What is clearer now is that much of the performative statement-making around the AlphaGo system which enacted it as a beyond-human, rupturing force able to make ultimate arbitrations came about not only because of qualities it exhibited, but because of the resources it was composed of as well. Technologies that require a high control of resources lend themselves towards an authoritarian or limited range of control (Winner, 1980). Contrasting the LeelaZero system to AlphaGo, much of what facilitated the separated status of the latter seems absent. Through an open and crowdsourced enactment, LeelaZero not only achieves what the AlphaGo system did, it is inherently more accessible. As a result, while still seen as playing beyond human skills, LeelaZero is not marked and therefore realized as wholly separated from humans.

Much of the distance and allure of the AlphaGo system came about from the guarded quality and approach in which DeepMind demonstrated – but never made it available – to the world. With the accessibility that the LeelaZero system provides, as we’ve already seen, players are now able to incorporate the insights that go-playing AI systems help produce into their own training and gameplay with better understanding of the reasoning processes each system demonstrates (Shin et al., 2021). Even playing games online now includes an AI review available at the end, enabling opponents to review their matches immediately and even discuss their shared understandings if they want. Interpretability and explainability, key values which were absent with the AlphaGo system and added to much of its oracle/separated status, are present here. Players can continue to study the value of a variety of different moves and gameplay strategies with LeelaZero. Rather than being unattainable qualities of advanced AI systems in general, LeelaZero helps highlight these dimensions were missing because they were not priorities for DeepMind or its enrolled allies, and because access was never made possible with the AlphaGo system. Given access and reflection, human players show that superhuman go-playing AI systems are not as alien as we may perceive if we only attend to AlphaGo and official discourses around it. What this access has also revealed is that limits still exist for superhuman go AIs too. As a pre-print article released February 2023 shows, there are blind spots in superhuman go AI systems, such that playing moves which even human amateurs can respond to successfully unravels the current top-playing AI system (Wang et al., 2023). Not only are superhuman AIs not divinely incomprehensible, neither are they infallible.

The technological reconstitution/reconstruction of AlphaGo with LeelaZero highlights how the go community engaged in both inverting the supposed symbolism of AlphaGo, and how this resulted in the creation of a technological production process and artifact which negate and reverse the supposed implication of AlphaGo for the go-playing community – opening up whole new playstyles and opportunities for human players, rather than shifting into any stage of finality.

8. Conclusion

Subject to a substantial amount of attention, the media hype and attention around AlphaGo often influences scholarship on the system too. Moving from a critical position which recognizes the agency of impacted constituencies when facing an overwhelming force, I employed empirical materials from a multidimensional perspective to illustrate the engagement and resistance processes which go players developed after the announcement and various iterations of AlphaGo. Borrowing the conceptual language of Pfaffenberger’s (1992) *Technological Dramas*, I illustrated engagement and resistance as both “technological adjustment:” the myth-, context-, and artifact-altering practices players use in accommodating the AlphaGo system; as well as “technical reconstitution:” the conscious attempts to change the system, or in the case examined here, create an alternative system (which I term reconstruction). This empirical analysis adds a novel interpretation of the system, and identifies a wider repertoire of responses to AlphaGo than is currently present in the literature, while the focus on engagement by impacted constituencies can be of interest to wider audiences in critical data/algorithm studies and social studies of AI.

Attending to the analytic category of impacted constituencies opens up the space to recognize these social groups as possessing a novel kind of agency which thus far has been theorized in narrow ways as passive enchantment or disenchantment. Employing an asset-based understanding of impacted constituencies reveals resistive practices where system owner narratives and curated meanings are challenged and reshaped by the engagement of those facing consequences without avenues of direct redress available. While technical programming skills are valuable in this repertoire of action, so too are communal-based, and supportive practices which involve interactions among the impacted constituency, such as sharing data, collective meaning-making through dialogue, and sharing strategies which minimize harmful or deleterious effects of the AI system. While the stakes may appear relatively low in the case of

a go-playing AI system, the practices examined here can be analytically and empirically revealing of similar avenues of resistance and engagement in cases where serious harms are present. For example, in cases of so-called AI-enabled “predictive” policing, we can see a similar repertoire of action as impacted constituencies engage in counter-narrativizing, withholding data, and public education campaigns within communities to raise awareness of how new AI systems used by the police are transforming and intensifying harms (Coalition, 2021; Jefferson, 2020; Minocher & Randall, 2020). Moving beyond the predominant focus on the data going into these systems, and studies attempting to ameliorate expressed harms, studies of engagement with AI in domains such as these can challenge our own imaginaries of what is possible in response to the increasing digitalization and controlling of our social worlds. Overall, a focus on impacted constituencies and their engagement and resistances emphasizes the importance of a processual lens in assessing discourse around AI systems, and the need to include multiple perspectives grappling with or making claims for what the impact of an AI system is, will, or should be.

Acknowledgements

I'm grateful to the Haverford faculty writing group led by Paulina Ochoa Espejo for the group accountability, feedback, and support in completing this paper.

References

- Benjamin, R. (2016). *Catching Our Breath: Critical Race STS and the Carceral Imagination*. *Engaging Science, Technology, and Society*, 2(June), 145–156. <https://doi.org/10.17351/ests2016.70>
- Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.
- Biernacki, R. (2014). Humanist Interpretation Versus Coding Text Samples. *Qualitative Sociology*, 37(2), 173–188. <https://doi.org/10.1007/s11133-014-9277-9>
- Binder, W. (2021). AlphaGo's Deep Play: Technological Breakthrough as Social Drama. In *The Cultural Life of Machine Learning* (pp. 167–195). Springer International Publishing. https://doi.org/10.1007/978-3-030-56286-1_6
- Binder, W. (2022). Technology as (Dis-)Enchantment. AlphaGo and the Meaning-Making of Artificial Intelligence. *Cultural Sociology*. <https://doi.org/10.1177/17499755221138720>
- Bory, P. (2019). Deep new: The shifting narratives of artificial intelligence from Deep Blue to AlphaGo. *Convergence*, 25(4), 627–642. <https://doi.org/10.1177/1354856519829679>
- Broussard, M. (2019). *Artificial Unintelligence: How Computers Misunderstand the World*. The MIT Press.
- Brown, P. (1992). Popular Epidemiology and Toxic Waste Contamination: Lay and Professional Ways of Knowing. *Journal of Health and Social Behavior*, 33(3), 267–281.
- Calo, R., & Citron, D. K. (2021). The Automated Administrative State: A Crisis of Legitimacy. *Emory Law Journal*, 70(4), 797–845. <http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=149980004&site=ehost-live>
- Campolo, A., & Crawford, K. (2020). Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society*, 6, 1. <https://doi.org/10.17351/ests2020.277>
- Clarke, A. E., & Charmaz, K. (2019). Grounded Theory and Situational Analysis. In *SAGE Research Methods Foundations*. <https://doi.org/10.4135/9781526421036825838>
- Coalition, S. L. S. (2021). Data-driven policing: Abolition requires a culture of resistance. *The Abolitionist*, 11–12.
- Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press. <https://doi.org/10.7551/mitpress/12255.003.0004>
- Curran, N. M., Sun, J., & Hong, J. W. (2020). Anthropomorphizing AlphaGo: a content analysis of the framing of Google DeepMind's AlphaGo in the Chinese and American press. *AI and Society*, 35(3), 727–735. <https://doi.org/10.1007/s00146-019-00908-9>
- Dayan, D., & Katz, E. (1994). *Media Events: The Live Broadcasting of History*. Harvard University Press.
- D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. The MIT Press.
- Egri-Nagy, A., & Törmänen, A. (2020). The game is not over yet—go in the post-alphago era. *Philosophies*, 5(4), 1–16. <https://doi.org/10.3390/philosophies5040037>
- Ensmenger, N. (2011). Is chess the drosophila of artificial intelligence? A social history of an algorithm. *Social Studies of Science*, 42(1), 5–30. <https://doi.org/10.1177/0306312711424596>
- Epstein, S. (1995). The Construction of Lay Expertise: AIDS Activism and the Forging of Credibility in the Reform of Clinical Trials. *Science, Technology & Human Values*, 20(4), 408–437. <https://doi.org/10.1177/016224399502000402>
- Feinstein, N. (2011). Salvaging science literacy. *Science Education*, 95(1), 168–185. <https://doi.org/10.1002/sce.20414>

- Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2), 219–245. <https://doi.org/10.1177/1077800405284363>
- Gibney, E. (2016). Google AI algorithm masters ancient game of Go. *Nature*, 529(7587), 445–446. <https://doi.org/10.1038/529445a>
- Guly, Christopher. (1998, December 31). "Deep Blue Turns Its Computing Talents to Worldly Pursuits." *Financial Post*.
- Hamilton, S. N. (2000). The Last Chess Game: Computers, Media Events, and the Production of Spectacular Intelligence. *Canadian Review of American Studies*, 30(3), 339–360. <https://doi.org/10.3138/cras-s030-03-05>
- Humphreys, L. (2005). Reframing social groups, closure, and stabilization in the social construction of technology. *Social Epistemology*, 19(2–3), 231–253. <https://doi.org/10.1080/02691720500145449>
- Jefferson, B. (2020). *Digitize and Punish: Racial Criminalization in the Digital Age*. University of Minnesota Press.
- Kline, R., & Pinch, T. (1996). Users as agents of technological change: The social construction of the automobile in the rural United States. *Technology and Culture*, 37(4), 763–795. <https://doi.org/10.2307/3107097>
- Law, J. (1986). On the Methods of Long Distance Control: Vessels, Navigation, and the Portuguese Route to India. In *Power, Action and Belief: A New Sociology of Knowledge?* (pp. 234–263).
- Leela Zero. (n.d.). <https://zero.sjeng.org/>
- Marris, C. (2015). The Construction of Imaginaries of the Public as a Threat to Synthetic Biology. *Science as Culture*, 24(1), 83–98. <https://doi.org/10.1080/09505431.2014.986320>
- Michael, M. (2009). Publics performing publics: Of PiGs, PiPs and politics. *Public Understanding of Science*, 18(5), 617–631. <https://doi.org/10.1177/0963662508098581>
- Minocher, X. (2018). Online consumer activism: Challenging companies with Change.org. *New Media & Society*, 1–19. <https://doi.org/10.1177/1461444818803373>
- Minocher, X., & Randall, C. (2020). Predictable policing: New technology, old bias, and future resistance in big data surveillance. *Convergence: The International Journal of Research into New Media Technologies*, 1–17. <https://doi.org/10.1177/1354856520933838>
- Moss, E., & Schüür, F. (2018). How Modes of Myth-Making Affect the Particulars of DS/ML Adoption in Industry. *Ethnographic Praxis in Industry Conference Proceedings*, 2018(1), 264–280. <https://doi.org/10.1111/1559-8918.2018.01207>
- Natale, S., Bory, P., & Balbi, G. (2019). The rise of corporational determinism: digital media corporations and narratives of media change. *Critical Studies in Media Communication*, 36(4), 323–338. <https://doi.org/10.1080/15295036.2019.1632469>
- Nguyen, C. T. (2020). *Games: Agency as Art*. Oxford University Press.
- Oh, C., Lee, T., Kim, Y., Park, S. H., Kwon, S., & Suh, B. (2017). Us vs. Them: Understanding artificial intelligence technophobia over the Google DeepMind Challenge Match. *Conference on Human Factors in Computing Systems - Proceedings*, 2017-May, 2523–2534. <https://doi.org/10.1145/3025453.3025539>
- Olazaran, M. (1996). A Sociological Study of the Official History of the Perceptrons Controversy. *Social Studies of Science*, 26(3), 611–659. <https://doi.org/10.1177/030631296026003005>
- Oudshoorn, N., Rommes, E., & Stienstra, M. (2004). Configuring the User as Everybody: Gender and Design Cultures in Information and Communication Technologies. *Science Technology and Human Values*, 29(1), 30–63. <https://doi.org/10.1177/0162243903259190>
- Parthasarathy, S. (2010). Breaking the expertise barrier: Understanding activist strategies in science and technology policy domains. *Science and Public Policy*, 37(5), 355–367. <https://doi.org/10.3152/030234210X501180>
- Pfaffenberger, B. (1992). Technological Dramas. *Science, Technology & Human Values*, 17(3), 282–312. <https://doi.org/10.1177/016224399201700302>
- Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, 14(3), 399–441. <https://doi.org/10.1177/030631284014003004>
- Selber, S. A. (2004). Technological dramas: A meta-discourse heuristic for critical literacy. *Computers and Composition*, 21(2), 171–196. <https://doi.org/10.1016/j.compcom.2004.04.001>
- Shin, M., Kim, J., & Kim, M. (2021). Human Learning from Artificial Intelligence: Evidence from Human Go Players' Decisions after AlphaGo. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1795–1801.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484. <https://doi.org/10.1038/nature16961>
- Sormani, P. (2023). Interfacing AlphaGo: Embodied play, object agency, and algorithmic drama. *Social Studies of Science*. <https://doi.org/10.1177/03063127231191284>
- Tavory, I., & Timmermans, S. (2014). *Abductive Analysis: Theorizing Qualitative Research*. University of Chicago Press.
- Taylor, P. J. (1995). Building on construction: An exploration of heterogeneous constructionism, using an analogy from psychology and a sketch from socio-economic modeling. *Perspectives on Science*, 3(1), 66–98.
- Velkova, J., & Kaun, A. (2019). Algorithmic resistance: media practices and the politics of repair. *Information Communication and Society*. <https://doi.org/10.1080/1369118X.2019.1657162>

- Wang, T. T., Gleave, A., Tseng, T., Belrose, N., Pelrine, K., Miller, J., Dennis, M. D., Duan, Y., Pogrebniak, V., Levine, S., & Russell, S. (2023). Adversarial Policies Beat Superhuman Go AIs. *ArXiv Preprint*, 1–58. <http://arxiv.org/abs/2211.00241>
- Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, 109(1), 121–136. <https://doi.org/10.4324/9781315259697-21>
- Woodhouse, E. J. (2005). (Re)Constructing Technological Society by Taking Social Construction Even More Seriously. *Social Epistemology*, 19(2–3), 199–223. <https://doi.org/10.1080/02691720500145365>
- Wynne, B. (1992). Misunderstood misunderstanding: Social identities and public uptake of science. *Public Understanding of Science*, 1(3), 281–304. <https://doi.org/10.1088/0963-6625/1/3/004>
- Ziewitz, M. (2016). Governing Algorithms: Myth, Mess, and Methods. *Science Technology and Human Values*, 41(1), 3–16. <https://doi.org/10.1177/0162243915608948>

AI policymaking as drama

Stages, roles, and ghosts in AI governance in the United Kingdom and Canada

Alison Powell¹ and Fenwick McKelvey²

¹ London School of Economics, United Kingdom

² Concordia University, Canada

✉ fenwick.mckelvey@concordia.ca

Abstract

As two researchers faced with the prospect of still more knowledge mobilisation, and still more consultation, our manuscript critically reflects on strategies for engaging with consultations as critical questions in critical AI studies. Our intervention reflects on the often-ambivalent roles of researchers and ‘experts’ in the production, contestation, and transformation of consultations and the publicities therein concerning AI. Although ‘AI’ is increasingly becoming a marketing term, there are still substantive strategic efforts toward developing AI industries. These policy consultations do open opportunities for experts like the authors to contribute to public discourse and policy practice on AI. Regardless, in the process of negotiating and developing around these initiatives, a range of dominant publicities emerge, including inevitability and hype. We draw on our experiences contributing to AI policy-making processes in two Global North countries. Resurfacing long-standing critical questions about participation in policymaking, our manuscript reflects on the possibilities of critical scholarship faced with the uncertainty in the rhetoric of democracy and public engagement.

Keywords: policy; drama; AI governance; Canada, United Kingdom; critical policy studies; hauntology

1. Introduction

The Honorable Minister of Innovation, Science and Industry François-Philippe Champagne (2023) rose to introduce Canada’s new AI regulation, the Artificial Intelligence and Data Act, before a Canadian parliamentary committee. Champagne extolled its virtues and the process: “I think the bill shows that we have listened to Canadians and parliamentarians, as we have done for other legislation. I gather there were more than 300 meetings or consultations with individuals who provided input into the process.” Many questioned when and where those meetings happened. Critical technologist Andrew Clement finally received the list – only to discover that the consultations mostly happened with industry (and after the Bill’s release) (Castaldo, 2023). One of the authors appeared on the list of consultees – but for discussions that took place after the Act had launched. So much for being an influential expert. Instead, what had been perceived as a friendly meeting became part of a spectacular number, part of the opening act in the Act’s dramatic passage. This scene is one of many that have caused to your authors to wonder:

How can two policy scholars make sense of AI governance, of technology policy, if the consultations seem so staged?

What began in our self-reflection has become a deeper attempt to understand the particular complexity of AI as technology governance. Your authors started to reflect on the roles afforded to academics and critical technology policy scholars asked to participate in the emerging field of AI governance, when our expertise in policymaking also rests on our capacity to be continuously up-to-date with emerging technologies that are often being promoted in ways that mystify, rather than clarify their social impact (Powell, forthcoming). Our collaboration – emerging from shared debriefs about being played-out or outplayed in what-can-seem-like consultation theatre – turned toward a broader reflection on what AI governance is as a certain mode of publicity, of a way of public-sense making about a new technology. Rather than focusing on matters of framing or problematization (Bacchi, 2009; van Hulst & Yanow, 2016), we have instead considered the dramatic function of policy debates to specifically question the roles and opportunities provided to scholars participating in these debates.

Our contribution to the special issue is a reframing of the terms of the AI policy debate from two technology policy scholars, on two sides of the Atlantic contending with wave after wave of AI controversy and consultation. Our tempest is the drama that has become AI governance. Our reflections take place amidst growing demand for consultations to legitimate artificial intelligence and specifically generative AI by governments and industry (cf. Palmås & Surber, 2022). Our reflection concerns AI's policy publicities as well as our sense of a growing ambivalence in the face of an expectation to play one's part in the legitimation of technoscientific capitalism (Birch, 2017; Nagy & Neff, 2024). Our article unfolds as follows: we first theorize our position on policy making as drama, then discuss how we can use this to understand current AI policy research agendas. Next, using examples from AI policy-making dramas in both the UK and Canada, we discuss how scenes are set through reference to different issues, giving benefit to some players over others. We then outline the roles that academics are invited to play as consultants, 'good academics,' maybe even as policy operators. This offers a new way to understand AI governance by attending to the drama as process, the staging of scenes that coordinate participation, the roles afforded to academics, and, oh yes, the ghosts. The stages where policy dramas play out can include ghosts - unacknowledged presences of past and active exclusions. We conclude by reflecting on how our actions as policy actors can be haunted by our past efforts, and what we might gain (or lose) from ghosting the process.

2. Understanding AI policy as drama

We theorize (along with many others) that policy making, like so many human endeavours, is dramatic. There are policy actors, there is a stage, and there are different scenes that unfold as the narrative progresses. We find ourselves on these stages from time to time. Different scenarios form set piece scenes in the narrative: public consultations where the chorus is invited to recite lines drawn from the rough-and-tumble world of experience, or closed, Chatham-House-Rules meetings where side conversations in the corridors reveal or obscure the potential for agreements, misalignments, or capitulations. Sometimes, metaphorically speaking, someone is concealed behind an arras. If occasionally the drama bursts on to screens, more often the stages seem mundane.

Drama is an antagonist for conventional policy scholars. Harold Lasswell co-constructed the policy sciences as expert-led, scientific project for governments and academics to govern societies (Turnbull, 2008). Merely suggesting that policy making might involve power plays breaks character with this part of the field's foundations. Certainly, we risk tarnishing our reputation as competent policy scholars by emphasizing the drama, but you can only watch so many AI executives act out their version of Oppenheimer-like regret before questioning the policy sciences. Instead, we observe how Schön and Rein (1994) argue that making better policy does not require always making 'better' rational decisions, but instead involves the policy participant having a more refined awareness of their own position in the

drama. Stone (1997) argues that policy making always involves conflicts over ideas, which are enacted through staged arguments. These orientations towards the dramatic, along with the acceptance of the diverse and often competing roles that play out across multiple scenes, help to clarify what policy making - and policy conflicts - are doing.

Our approach then differs from much of the other policy literature on AI in considering AI governances as a dramatic process. Much of the existing literature focuses on the consequences of policies about AI (Acemoglu, 2021; Angwin et al., 2016; Buolamwini & Gebru, 2018). On one hand, research has focused on AI governance as addressing existential challenges (not risk) where research aims to define key problems (e.g., explainability, traceability, or bias) that necessitate governance reforms (Büthe et al., 2022; Gasser & Almeida, 2017; Knight et al., 2024). At the most abstract level, AI has prompted a fruitful debate over the nature of contemporary governance and governmentality (Amoore, 2022; Fourcade & Gordon, 2020; Yeung, 2018). On the other, there is a growing turn toward understanding the discourses of national AI strategies as evidence of empirical policy trends. Studies range from comparative evaluation to more critical work describing the underpinning imaginaries that legitimate these policies (Bareis & Katzenbach, 2021; Hansen, 2021; Vicente & Dias-Trindade, 2021). We find inspiration in both joining to a class of spectators wondering how long AI government can maintain the status quo while claiming AI is revolutionary while also attending to the everyday efforts of AI governance. Amidst these two trends, our approach joins a growing interdisciplinary interest in the practical roles of AI governance (Sætra et al., 2022; Widder, 2024).

In this paper, we position our lived experiences and academic engagement in relation to what might be called a performative turn in critical policy studies (Clarke et al., 2015; Lea, 2020; Lipsky, 2010). John Clarke, Dave Bainton, Noémi Lendvai, and Paul Stubbs inspired us in their writing on policy as: “a way of imagining the world as an object of intervention; as a way of enrolling subjects into a process of acting; and as a practice that seeks to produce effects, including the act of ‘taking the politics out of things’” (2015, p. 34). This allows policy to be understood as something that itself moves or contains movement and dynamism, “unlocking the narrow confines within which ‘policy’ is conventionally enclosed - the rational, technocratic world” (2015, p. 33). Through this we can explore how contradictory entities tussle between and within themselves to create ideas that position and discipline different groups of people in different ways. We aim to provide a language to understand AI governances as lived drama, beginning with, an assessment of how policy is staged.

2.1 Breaking the fourth estate / wall: Policy, stages and scenes

“Mr. ChatGPT goes to Washington” read the headline of the CNN title about Sam Altman’s testimony before Congress (Fung, 2023a). Mr. Altman showed up before the television cameras to play the role of a good CEO. Why not compare the testimony to a movie? Altman had to wear a suit – a costume his now-jailed contemporary Sam Bankman-Fried only started to put on when he, too, went on trial. Altman exceeded expectations in his role. One glowing review of Altman’s performance reads “OpenAI CEO Sam Altman seems to have achieved in a matter of hours what other tech execs have been struggling to do for years: He charmed the socks off Congress” (Fung, 2023b). What a way to win over the audience of regulators. Altman’s success in staging gives us a good reason to understand AI policy as drama. The suit helped put him the role as an ethical technoscientific capitalist. Regardless of what might have transpired inside OpenAI, Altman was obliged to enact a responsible proxy for the firm. Altman’s successful performance – one that a critic might argue displaced attention from deep concerns about the direction of the industry – is presently what has us wondering about the extent to which policy success is in the staging, and how this serves to conform to expectations of the performance – and thereby to influence the outcome.

If critical policy studies have a public role in responding to AI (Paul, 2022), then we encourage a reflection on how policy consultations are staged – often as media events – and to what extent this staging

displaces attention that might otherwise focus on the “epistemic forgeries” (Katz, 2020), at the heart of AI. These forgeries include assumptions that AI innovation is inevitable – and that it can be made unproblematic or even ‘safe’ if only policy processes can be properly focused. The staging of policy dramas reflects this effort to contain and constrain policy making. Consider some of the settings such as the computer science conferences run by the ACM and the closed meetings hosted by technology companies, venerable Institutions, or government agencies where the drama unfolds away from the public eye. Of course, some actors have the capacity to move across various settings, shaping how the dramas are staged: OpenAI, Meta, Alphabet, Microsoft. If you watch long enough, you might see actors change into corporate costumes after playing other roles - perhaps in research, or within regulatory bodies – revolving doors operate on these stages too (Bates, 2013).

The AI policy making stage has also been growing, and the players are striding out into a runway cutting all the way through the theatre. The fourth wall is breaking. Members of the audience are leaping up and declaring interests in being part of this drama, too, proclaiming an interest in the proceedings. But to be sure, this is not Brechtian theatre where the audience can play along. A friction endures between the construction or instantiation of a public and policy formation (Marres, 2005). There is growing attention to the “consultation theatre” around AI, a charade where the consultation serves more as a chance for governments to perform their role as stewards of public opinion and showcase their ability to respond, but not actually to affect to shape policy (M. Jones & McKelvey, 2024). Not being able to look away might be the best defense of democracy these days when the public’s only motivation is to watch the drama as a spectator.

Policy drama might also be an effect of how the mass media plays expectations on the policy process, what Maarten Hajer (2009) identifies as a ‘mediatized’ policy making. From this perspective, the cinematic aspect of Altman’s US Congress appearance is not accidental. Steven Coleman (2013) describes how politics construct drama, and how cultural performance applies not only to the micro-interactions of interpersonal relationships but also to macro-political organizations like television. People justify their interactions through drama, or what Isabela and Norman Fairclough (2013) call the ‘practical reasoning’ behind these processes. A reasoning that, we find elsewhere, journalism has largely failed to provide, getting publics to worry about existential risks and taking the promise of large-language models as the next step to general artificial intelligence at face value (McKelvey & Roberge, 2023). But we are foreshadowing too much; let us show how AI policy scenes are set, and how the change of these scenes influences which voices are heard on stage.

2.2 From ethics to existential risk to safety - the UK policy scenes 2016-2024

One of your authors has watched many of these scene changes in the UK over the past decade. Concern about ethics and AI started as a problem for the ‘good society’ (Cath, 2018; Floridi, 2014). The policy scene expected players to describe AI’s ethics and values (Jobin et al., 2019; Mittelstadt et al., 2016). Action in the mid-2010s focused on ethical principle statements for AI (European Union, 2019; UK Government, 2019), and subsequent discussions of how to put them into practice (Stark & Hoffmann, 2019). For a brief period, corporate actors played along, issuing statements and constituting company ethics boards (Taylor & Dencik, 2020). By 2020, however, AI harms were accruing regardless of these ethics statements (Morley et al., 2023).

The academic aspects of the policy scene started to consolidate into discussions drawing on disciplinary expertise, for example, involving technical concerns about model design or auditing capacity (Sloane, Moss, Awomolo, et al., 2022; Sloane, Moss, & Chowdhury, 2022) – changing the language and behavior of the scene in ways that sometimes left players like your authors in the wings (Gansky & McDonald, 2022).

Then the scene changed. Different stories tried to nuance the discussion of ethics, expanding it to cover various areas of practice, and then to attempt to connect these suggested areas of practice to regulatory

capacities (Sætra et al., 2022). In the UK, these efforts drew from traditions of technology policy-making focused on responsibility (Owen et al., 2013; Stilgoe, 2018, 2019) and engineered-in trustworthiness (Naiseh et al., 2022; Schneiders et al., 2023).

These scenes structured what was focused on, by whom, and at what time. A mismatch with the unfolding scene might create confusion or result in being moved off stage as we have experienced. One author was part of the JUST AI project, a pilot hosted by then-recently-founded Ada Lovelace Institute. JUST AI employed feminist lab methods to convene conversations and hold space for dissent and foregrounded the inclusion of arts and artistic methods in reframing how the conversations about AI ethics were occurring – away from the statements of principle and corporate boards and in a collaborative, collegial, and slow-moving manner. JUST AI’s networked approach was a backdrop against which to choreograph the entries of new actors whose script and position could generate research, conversation, and action on racial justice, reparability, and environmental sustainability as aspects of AI ethics work. The scenes in play in the early 2020s UK ethics drama complicated this. A JUST AI Fellowship focused on racial justice in AI supported artists and creative researchers interested in the use of AI in border technologies, the policy gaps relating to differential impacts of AI on racialized populations, and how the white supremacist origins of AI might be re-imagined (Chander & Nation, 2021), which challenged some of the more reductive discussions about bias mitigation (Buolamwini & Gebru, 2018; Gong et al., 2020). The unfolding conversations on these topics joined other working group discussions of the environmental ethics of AI (including its reparability), and the a disability-led discussion of the potential for both rights of access to AI and rights to its refusal (Hickman & Serlin, 2018; Newman-Griffis et al., 2023). These ethical conversations were bracing, contentious, and sometimes difficult to place within the policy scene. JUST AI events convened wide public audiences and influential academic participants (JUST-AI, 2023), but the project’s focus on systemic injustice and large-scale environmental impact sometimes felt out of synch with other work at Ada Lovelace that engaged directly in the policy scenes attracting attention at the time.

The scene changed again; the revolving door whirled, and by 2023 the focus was on risk and safety, rather than ethics. This was solidified in the European Union’s AI Act, which rested on an assumption about managing and regulating riskier applications of AI. But this script was influenced by actors with particular interests (Hendrycks & Mazeika, 2022; Turchin & Denkenberger, 2020) and increased attention to the capacities of Large Language Models, especially ChatGPT. A new set of actors introduced the idea of existential risks to human society that could result from significant technical developments in AI. Led by Nick Bostrom in Oxford and backed by significant funding from the technology industry, the ‘long-termist’ approach focused on the existential risks to the entire human race of unimpeded, speculative AI development (Gebru & Torres, 2024). By late-2023 the language had shifted from existential risk to safety – with the initial discussion still focused on large-scale threats and ‘doomsday’ scenarios now also accompanied by concerns about the development of multi-purpose or ‘foundation’ models in response to the commercialization of ChatGPT. In the dramatic and media-friendly AI Safety Summit convened by the UK Prime Minister in November 2023 highly visible actors called for the testing of AI ‘foundation’ models and the management of AI “for good” in a ‘pro-innovation’ framework (Hawes & Hall, 2023). The structure of the AI Safety Summit as an invitation-only meeting of many industry representatives, a handful of high-level politicians and selected academics (not your authors), foregrounded a cheery, media-friendly narrative of ‘safety’. Very few of the UK’s diverse range of civil society stakeholders were invited. A parallel set of events branded as the AI Fringe (2023) was organised by a research consultancy and bundled together sector-specific discussions of including AI and work, access rights, and public accountability.

The resulting “Bletchley Declaration” focused on the safety of ‘foundation models’ and called for the establishment of AI safety institutions – but might have had negligible impact on the UK’s current AI regulatory frame, which focuses on principles of safety, transparency and explainability, fairness and accountability, and contestability and redress (UK Government, 2024). Results from the AI Fringe have

included the TUC's AI Regulation and Employment Right Bill (Allen & Masters, 2024). Globally, more attention has been paid to the environmental impact of the use of AI systems, sometimes as an aspect of safety (Falk & van Wynsberghe, 2023) – demonstrating how a shift in the scene can bring an idea in from the wings.

These shifting scenes and focal points have required participants in policy dramas to develop a kind of improvisational agility, moving between scenes along with actors whose positions might also be changing. In the case of JUST AI, for example, the project's funding and institutional support rested on a partnership between the Ada Lovelace Institute, its background funder the Nuffield Foundation, and the UK's Arts and Humanities Research Council. The think-tank, the funders and the project all maneuvered across a rapidly changing policy scene, where reflections about ethics and the creation of new ways of networking across disciplines ceded space to direct interventions in government policy and technically-driven discussions about audit or features design (see Groves et al., 2024). JUST AI's contributions, including its Fellowship and commissioning of creative work, advanced ideas such as environmental sustainability that were not, at the time of the project, direct objects of policy development, although they have subsequently emerged as such. However, efficacy in policy making is often judged against experiences in the present rather than potential in the future, especially in a reactive context where researcher and other experts are called upon to continue to define or defend the 'public interest'. JUST AI's focus on environmental sustainability and repairability took some time to become salient within UK policy drama. To date, its work on rights, access, and refusal of AI systems from the perspective of critical disability studies has not yet become so. This is testament to what we've explored through this section – the ways that policy ideas, practices and potentials are captured within shifting scenes. Of course, within each of these scenes, scholars like us are called on to perform specific roles.

3. Some tentative roles for aspiring AI policy scholars

Your authors have been called to stage to perform as expert witnesses in Canada and the United Kingdom on a range of technology-influenced social issues, over the past decade and more. Now though, the intensity of AI consultations – this moment when governments and corporations are trying to advance AI's acceptability – have us questioning what it means to play a role to participate in AI policymaking.

Roles are both individual and collective. The role “refers to a cluster of behaviors and attitudes that are thought to belong together, so that an individual is viewed as acting consistently when performing the various components of a single role and inconsistently when failing to do so.” (R. H. Turner, 2001, p. 233). Within these roles, Peter Burke claims that actors (or agents), exercise their agency in ways that are proper to the scenes they are participating in. “It is a principle of drama that the nature of acts and agents should be consistent with the nature of the scene” (Burke, 1945, p. 3). As your authors don and doff various costumes and move between various roles, we keep in mind these insights. Language matters in these settings, as Coleman identifies, but performance does too. Performing is “a culturally pragmatic strategy for the production of shared meaning” (Coleman, 2013, p. 330). Policy making, especially technology policy making, contends with both increasingly mediatized aspects of decision making (where strategic media visibility can mean leveraging policy positions) and with the influence of entrepreneurial innovation contexts (see Irani, 2015), which have also impacted the roles that professors and university researchers occupy. This generates conflicts between roles that different actors can play, and the extent and quality of their agency.

Based on our shared experience, we offer a few possible roles a policy scholar might play. In the next section, we draw on our experiences and observations as experts in the spaces to identify where academics, researchers, and intellectuals – domain experts, in other words – are invited and expected to act in AI policy-making dramas. We also explore the space of ambivalence or haunting that can accompany this performance of expertise and the strange experiences that can result when our actions

seem out of line with roles, are haunted by what came before, or suggest potential futures that jar with the scenes at play.

3.1 *The Consultant / Hired Gun / The Ronan*

The first role is that of an intellectual worker as a consultant-for-hire. Getting paid is not the issue (participation should always be fairly compensated); rather, the concern relates to the point at which expertise becomes sponsored research (see Goldenfein & Mann, 2023). The role of the consultant acknowledges that much of academia maintains relationships with corporate partners that might aid or support broader lobbying agendas. Academics have been accused of being too close to the AI industry, or what Meredith Whittaker describes as “the steep cost of capture” where “the tech industry’s dominance in AI research and knowledge production puts critical researchers and advocates within, and beyond, academia in a treacherous position” (2021, pp. 51–52). The issue has become particularly pronounced in regards to sponsorship from artificial intelligence firms, where the Washington Post finds that 6 out of 10 AI researchers with disclosures of funding come from AI (Menn & Nix, 2023). Dependence does not, necessarily, determine participation in policy, but more structurally, these findings establish a synergy between corporate and academic institutions that may translate into tacit support. Luke Stark at Western University made an important statement in 2021 when he turned down \$60,000 of Google funding after Google fired many of its AI ethics teams (Metz, 2021). Stark’s decision illustrates the rarity and potential of refusal, highlighting the fact that more commonly decisions are made that involve accepting funding and then later navigating fraught relationships.

3.2 *The Good Academic*

A proper academic would never have written this section. They would review submissions to policy consultations to identify emergent frames, ideally using double-blind coding, and then present a weighted summary of positions and key clusters. We did not do that. Instead, we quickly popped backstage and glanced over the script to see what scene we were playing now based on a story in the media. We saw a ghost in the corner, a memory of how this played out last time. We know how to play the “good academic” we are just not doing it right now.

One role for “the good academic” is to act as an independent voice, playing a reliable role as a good policy actor. This role is part of the foundation of policy studies, which Laswell argues are “concerned with knowledge of and in the decision processes of the public and civic order” (1971, p. 1). Lasswell called for these positions to be defended, to be independent, and capable of being objective. This means that the ‘good academic’ is meant to be doing the work of doing research, in line with the expectations and norms of their field. The ‘good academic’ role differs from an on-demand consultant because their obligations come from the field, empiricism, and objectivity, rather than the contract.

A final consideration and one that is a looming challenge for objective AI research in general, is how “good academics” perform objectivity amid debates over methodology and critiques of their approaches. Presenting technical expertise in policy-making settings can put great strain on the role. An example of such a strain came from a performance one of the authors made during policy debates about Net Neutrality – a precursor to AI debates in requiring proof of the presence of bias in an infrastructural technology (McKelvey, 2018). Fenwick has an embarrassing tale of working with computer scientist Dave Choffnes to prove that one of Canada’s mobile service providers had been modifying some of its telecommunication equipment to prioritize certain applications over others (Li et al., 2016). Done at the last minute, this transcript is the only record remaining of the brief exchange:

103 McKelvey: I understand some of the – my argument here will introduce new evidence and I’m wondering if I have permission to introduce this evidence in support of our submissions earlier?

104 THE CHAIRPERSON: Why were you unable to provide this evidence earlier?

105 McKelvey: It's a great question. Working with a – by any serendipity honestly, I'm working in a new project working in measurement in Canada trying to understand and apply new internet measurement techniques to understand the operations of systems such as differentiate service provisions.

106 And it was only in early September that I knew of a methodology that is going to be presented at the internet measurement conference this November 2016 that I was – found out about it and it was basically over the past last month that I was able to actually deploy and use this measurement and try to support arguments we've made previously as part of our submission.

107 THE CHAIRPERSON: Okay. Well, we'll take your arguments under advisement but go ahead and ---

108 McKelvey: Okay.

109 THE CHAIRPERSON: But that doesn't mean ultimately we will necessarily admit it on the record.

With the evidence presented, questions turned to questions of telecommunication policy, and the infrastructure operator rejected Fenwick's assertion. Another opportunity for leveraging evidence in policymaking passed. The stage did not want to discuss academic methods, now was the time for deliberation of the law.

The example provides a warning to policy scholars of the limits of the "good academic" role here is not likely to be consensus on what counts as safe or risky AI, and no specific piece of evidence is going to be accepted as objective. Indeed, one strategy in climate science communication is for lobbyists to introduce facts as controversial, or debatable - just to raise enough doubt to thwart the efforts of the policy sciences (Hess, 2014). Fenwick's story should act as a corrective for any "good academic" hoping that some data and a novel method will expose bias and convince regulators. Instead, the method was not legible, indeed the whole project of offering "objective" Internet Measurement has become something of a hobby project since its only large Internet Service Providers willing to invest in the methods – maybe a warning to who is going to fund and legitimate the next AI audit?

3.3 The Policy Operator

The strains of the 'university in ruins' (Readings, 1997) drive the emergence of the policy operator, our third role. In this mode "the concept of scholarship as cultural production . . . is articulated through ideas about economically impactful deliverables, including a narrowly defined set of public engagement and outreach activities" (Luka et al., 2015, p. 177) . This role can involve working for hire as contractor, or alternatively an embrace of these norms to become something of a policy entrepreneur, speculating on the potential to develop new spaces of intervention. This can happen when "visibility and self-promotion naturally accompany a corporatized culture, requiring deliverables and impact to stand in for a broader conception of scholarship as cultural production" (Luka et al., 2015, p. 191). What could be more measurable than policy impact? The policy operator combines the pressures of the neoliberal university, the need for self-branding and greater visibility with the demand for good-faith participants in government policy. The mix, when successful, can be remarkable, producing a scholar able to manufacture a brand of authority, steer the policy agenda, and gain prominence through the new metrics of knowledge mobilisation. The policy operator then becomes a name who must be consulted. Your authors recognize our own ambivalence towards the policy operator that matches a sense of the position's power. We also recognize a concern over the demands to keep up, to be in the game, and to support or even benefit from a concentration of power in academia (noting its attendant inequalities, especially for those of us also tasked with managing students, programs, and departments alongside our performances in policymaking).

Canada's AI policy has been driven by policy operators, whose positions can grow to shape domestic and international policy scenes. Recently we have witnessed a strange performance by two of its three "godfathers of AI". These figures had occupied an oversized role in Canada's media as evangelists (Dandurand et al., 2023). Geoffrey Hinton and Yoshua Benigo then broke with the third 'godfather' Yann LeCun to publicly worry over AI's existential risks. Their rise to celebrity can matched 2023's summer blockbuster, *Oppenheimer* – or what was called the "Oppenheimer moment" (Tharoor, 2023). Hinton and

Benigo filled the role of today's Oppenheimers for AI – a role that had them make frequent testimony about the risks of AI.

The Oppenheimer moment helped legitimate Canada's AI policy agenda. Canada tabled its first AI legislation, the Artificial Intelligence and Data Act (AIDA), in Spring 2022. AIDA attracted little public attention until Benigo signed the Institute for the Future's Open Letter to Pause AI Experiments a year later, Spring 2023. As the letter attracted international attention, Bengio joined most of Canada's AI establishment to sign a second letter calling for the swift passage of AIDA as the kind of urgently needed regulation called for in the first Open Letter. The combination of letters became a mediated spectacle. The once-AI-evangelists offered a pious second act of regret and reflection, all to compel the bill's swift passage – a mix of personal drama and public pressure that seems a hallmark of the policy operator. Champagne (2023) in the speech mentioned above expressly mentions the first letter, claiming "You don't need to take it from me, but I would advise you to read the letter that was signed by Yoshua Bengio, and hundreds of people from around the world, warning us that we need to take action." The policy operator is one who knows how to command the drama to advance their policy agenda. (We say this as an observation, not commentary. Off-stage, the two letters reminded one author of what it means to not be a policy operator after signing a civil society letter calling for reforms to AIDA not mentioned in the Act's introduction.)

Our first three roles: the consultant, the good academic and the policy operator, create mechanisms to carry micropolitical sensibilities into the macropolitical, providing relatively well-specified opportunities for academics to leverage their sense of public accountability and influence. However, we notice that there are other, more ambivalent positions that involve divergent interpretations of the public interest. These other positions might engage what we refer to, following Tess Lea, as policy hauntology. This can occur through echoes of past policy paths not taken or through the framing of potential future policy objects. The unstable temporality of interventions can make these efforts seem haunted by ghosts of past or future.

4. A policy hauntology: Ghosts, ghosting and ambivalence

Critical policy theorist Tess Lea understands policy hauntology, as "the ways policies past and present are physically incorporated, having (insidiously or noisily), seeped into lives, affecting probable destinies and shaping overall circumstances, if not immediately then certainly as a powerful stimulus" (2020, p. 117). Jacques Derrida introduced hauntology at one of Marx's many wakes. Hauntology, to Derrida, meant, something else, "this element itself is neither living nor dead, present nor absent: it spectralizes" (1994, p. 63). Derrida uses the figure of the ghost as an object of hauntology "because it points toward a thinking of the event that necessarily exceeds a binary or dialectical logic, the logic that distinguishes or opposes effectivity or actuality" (1994, p. 78). The purpose here is to surface our shared collective frustration in having to perform AI as a new problem, as requiring "new" innovations, while haunted by past failed technology policies. Sometimes the whole drama seems just like another mediocre Hollywood update in the tired franchise of public interest technology regulation.

Technology policy is haunted by past technology. In Canada, this was illustrated by an attempt to AI and smartphones to automate COVID-19 contact tracing by applying AI models to tracking smartphone proximity. To work, the system needed a certain threshold of public adoption – a novel idea, but one haunted by Canada's ongoing challenges to extend smartphone and internet connectivity to all, and the mobile services regulatory frameworks that have resulted in unaffordable mobile services. The problem was not AI, but what happened before AI. Similarly, the difficulty in advancing issues of the environmental cost of AI in the UK was related to a policy context wherein the government at the time focused on extending oil and gas exploration rather than discussing climate commitments. Our sense of haunting has a lot to do with critical scholar Eve Tuck and C. Rae's interpretation of the function of horror in understanding cultures of justice. They write,

The difference between notions of justice popularized in US horror films and notions of justice in these examples of horror films from Japan is that in the former, the hauntings are positioned as undeserved, and the innocent hero must destroy the monster to put the world in balance again (though predictably, several of the hero's companions who are women or people of color will likely be sacrificed along the way). In the latter, because the depth of injustice that begat the monster or ghost is acknowledged, the hero does not think herself to be innocent, or try to achieve reconciliation or healing, only mercy, often in the form of passing on the debt. (Tuck & Ree, 2013, p. 642)

The haunting is a force and a demand - to acknowledge past failures, or to begin at a place of humility. At the very least it is a rejection of the hype. So, we welcome the horror of haunting - that danger that threatens to bring the curtains down - because the other option is to have the same old tropes performed once again.

Hauntology helps us understand what the silence threatens. Critic Mark Fisher (2012) discussed hauntology to understand the ways that innovation cannot escape prior social forms. The relationship between technology governance and industrial strategy has always been awkward. Technology in a Western context is often tied to economic growth and innovation economies. This alignment obfuscates human rights and social impacts or (perhaps at best) offers a calculus of trade-offs of how to manage innovation responsibly. These challenges predate artificial intelligence but resonate as soon as claims of a 'fourth industrial revolution' or a transformation in employment enter the discussion. That histories, those pasts, become what Tess Lea sees as a policy hauntology – “the deeply saturated effects of past policies.” (2020, p. 30).

Ghosts, however, are part of hauntology as observation, and hauntology as strategy. Writer and scholar Billy-Ray Belcourt wrote his *Poltergeist Manifesto* to acknowledge that “there are forms of life abandoned outside modernity's episteme whose expressivities surge with affects anomalous within the topography of settler colonialism” (2016, p. 24). The poltergeist is the “noisy ghost” that shows up when you build your house on the old graveyard. The noise speaks “speaking into existence an anti-subjectivity that emerges in the aftermath of death or murder” (2016, p. 26). A poltergeist is one ghost, a strategy for Belcourt, to find a queer indignity. Their ghost inspires us. The ghosts on stage are frankly unsettling to the whole performance. Climate is an excellent example. In a climate emergency, AI power demand is not only firing up old coal plants, it is diverting energy that might otherwise be used for housing (Simpson, 2024). The play now is to keep talking around the climate issue (Hogan, 2020) without disrupting anything fundamental. “Keep going, can't you, keep going!” Hamm yells in Beckett's play *Endgame*. Ghosts keep the play moving.

Maybe we can just up and ghost the process - not participate? Another kind of haunting is to be present in absence, or to settle into ambivalence. As Kearnes and Wynne argue, ambivalence can be interpreted as an “engaged – rather than passive – mode of relating to technological determinism” (2007, p. 131). Our efforts have generated such ambivalence. Perhaps we can see this kind of haunting in relation to how the JUST AI project engaged with AI policy issues, revealing temporal and conceptual ruptures in the policy through its attempt to reshape the models and practices through which AI policy was considered. JUST AI's lab models and slow discussions formed part of variously composed struggles against the automatic embedding of technologically-determinist understandings of innovation in technology policy, and the related expectation that social scientist policy scholars should be consistently up to date on specific technical risks. Working to frame issues before time and to resist constant extraction played at the ambivalent edge of participating and exiting.

To haunt could be to stay in the room but break the part by rejecting the frame, or to stay in the room and to operate rigorously in the interests of the most marginalised. Tess Lea foregrounds this approach when she describes the role of the ‘institutional killjoy’ who points out the future hauntings that the policy might produce, while also insisting on the integrity of the policy process. As Lea writes, “institutional kill-joys are those who, working hard to pull benefits from policy opportunities, insist that contracts be honoured, promises be kept, accounts be held, corruptions be resisted, projects have fidelity, and people

be respectfully supported” (2020, p. 14). The institutional killjoy may not be a good academic; but they are a good advocate and are often found acting stringently in the public interest.

A question raised by the killjoy’s role concerns how and in which ways these goals contradict the pace and scope of the policy drama as it unfolds. As such, this fourth role opens a question that may also haunt the policy actors as they enter the stage: what happens when policy processes become obscure, improvisational, or difficult to square with the previously considered notions of the ‘public interest’? What happens when an actor’s micro-sensibilities meet the messy meso-level of policy drama?

5. Ghosting the future: Ambivalent and generative outcomes

In determining whether our policy performances have been effective, we often examine the scene itself. Was there a Colombo moment, where the regulator was so swayed by evidence that we could yell, ‘aha!’? Unlikely. Did my performance sound confident? Did I read the room? What is more common is that the ghosts of the past haunt the present and prefigure the future. For drama to make sense we need to understand both the stage and what haunts it.

In this article we have consciously employed the terms of drama - staging, scenes, roles, and theatrical ghosts – to illuminate some of the complexities of academic participation in technology policy making, especially AI policy making. We have outlined how the staging of policy making can favour performances from already-powerful (often corporate) actors, and how shifting policy scenes and themes challenges the performance of expertise. We have briefly explored some of the roles that policy actors play in sustaining AI policy’s publicities, and the way that these roles are haunted by our past experiences and our anticipations of how developing technology in the public interest might, should, or could unfold. In our experience, we have both observed and occasionally played several of these roles, although we also observe how the intensification of AI policy-making’s publicities at present involves processes that produce erasures of policy topics and policy perspectives. The implications of these erasures include the potential undermining of epistemic justice and the exclusion of marginalised voices, or the thinning out of democratic oversight processes in favour of policy capture by louder (often commercial or corporate) voices. We contend that participation in the AI policy drama and its attendant publicities produces ambivalent positions, as represented by our ghosts on the stage of the drama. These ghosts remind us of what could have been as well as what still might become. Accepting this ambivalence, including its dynamism, improvisational and oppositional nature may be important to feed, nourish and develop the practice of the public interest, which lives in the messy meso-level between our own sense of justice and the macro structures of the world.

Our aims feel unfulfilled, and ambivalences remain. Yet scenes do change – perhaps not in the ways we expect. In Canada, the current AI bill seems dead in the water, likely to die with the end of Parliament all the while attention has shifted to AI safety (probably as part of Canada’s own performance on a global stage). For JUST AI, considerations of racial justice and ecological sustainability of AI systems initially slid out of view. This created a policy blind spot, since the results of the project were not easy to place within the shifting policy drama unfolding at the time. While the project’s efforts to bring strong standpoint epistemology and epistemic justice perspectives (see Fricker, 2007) into AI policy making may not have been remarked on the main stage, they may remain as ghosts to haunt future projects, or even to step into future policy scenes (Powell, forthcoming). One year after the project’s close, environmental sustainability has subsequently become a policy frame, while racial justice is addressed functionally in discussions of bias rather than through systemic efforts to consider the connections between innovation, exploitation, and systemic injustice. Disability-led design and the necessity for justified refusal of AI systems remain marginal policy topics that killjoys struggle to place in full view (Barron, 2024). Meanwhile, although large-scale strategic funding reframes the role of arts and humanities research as in service to, rather than critical of innovation (a recent strategic project “is dedicated to integrating Arts, Humanities and Social Science research more fully into the Responsible AI

ecosystem” (BRAID, 2023)), other funding calls focus on ‘a new model of research leadership and teamworking” and invites pilot experiments in organizing research more collaboratively (UKRI, 2014). Prototyping policy options may feel ambivalent, but the spectres it raises glimmer in the corners of future policy stages.

References

- Acemoglu, D. (2021). *Harms of AI* (Working Paper No. 29247). National Bureau of Economic Research. <https://doi.org/10.3386/w29247>
- AI Fringe. (2023). *AI Fringe*. <https://aifringe.org/>
- Allen, R., & Masters, D. (2024, April 18). *Artificial Intelligence (Regulation and Employment Rights) Bill*. <https://www.tuc.org.uk/research-analysis/reports/artificial-intelligence-regulation-and-employment-rights-bill>
- Amore, L. (2022). Machine learning political orders. *Review of International Studies*, 1–17. <https://doi.org/10.1017/S0260210522000031>
- Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). *Machine Bias*. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Bacchi, C. L. (2009). *Analysing Policy: What's the Problem Represented to Be?* Pearson.
- Bareis, J., & Katzenbach, C. (2021). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values*, 01622439211030007. <https://doi.org/10.1177/01622439211030007>
- Barron, D. (2024, June 3). *A radically inclusive approach to digital society*. Promising Trouble. <https://www.promisingtrouble.net/blog/a-radically-inclusive-approach-to-digital-society>
- Bates, J. (2013). The Domestication of Open Government Data Advocacy in the United Kingdom: A Neo-Gramscian Analysis. *Policy & Internet*, 5(1), 118–137. <https://doi.org/10.1002/poi3.25>
- Belcourt, B.-R. (2016). A poltergeist manifesto. *Feral Feminisms*, 6, 22–32.
- Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts.*: Hearing before the Standing Committee on Industry and Technology, House of Commons of Canada 1 (2023). <https://www.ourcommons.ca/DocumentViewer/en/44-1/INDU/meeting-86/evidence>
- Birch, K. (2017). Techno-economic Assumptions. *Science as Culture*, 26(4), 433–444. <https://doi.org/10.1080/09505431.2017.1377389>
- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Burke, K. (1945). *Grammar of motives*. Prentice-Hall. <http://archive.org/details/grammarofmotives1945burk>
- Büthe, T., Djeflal, C., Lütge, C., Maasen, S., & Ingersleben-Seip, N. von. (2022). Governing AI – attempting to herd cats? Introduction to the special issue on the Governance of Artificial Intelligence. *Journal of European Public Policy*, 29(11), 1721–1752. <https://doi.org/10.1080/13501763.2022.2126515>
- Castaldo, J. (2023, December 10). *Federal consultations on AI regulations heavily skewed toward businesses, industry groups, say critics*. The Globe and Mail. <https://www.theglobeandmail.com/business/article-canada-ai-law/>
- Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 20180080. <https://doi.org/10.1098/rsta.2018.0080>
- Chander, S. D., & Nation, S. (2021, June 24). *Who gets to write the future?* Ada Lovelace Institute. <https://www.adalovelaceinstitute.org/blog/who-write-future/>
- Clarke, J., Bainton, D., Lendvai, N., & Stubbs, P. (2015). *Making policy move: Towards a politics of translation and assemblage* (1st ed.). Bristol University Press; JSTOR. <https://doi.org/10.2307/j.ctt1t89fpb>
- Coleman, S. (2013). How to Make a Drama Out of a Crisis. *Political Studies Review*, 11(3), 328–335. <https://doi.org/10.1111/1478-9302.12024>
- Dandurand, G., McKelvey, F., & Roberge, J. (2023). Freezing out: Legacy media’s shaping of AI as a cold controversy. *Big Data & Society*, 10(2), 20539517231219242. <https://doi.org/10.1177/20539517231219242>
- Derrida, J. (1994). *Specters of Marx: The State of the Debt, the Work of Mourning, and the New International*. Psychology Press.
- European Union. (2019, April 8). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Fairclough, I., & Fairclough, N. (2013). *Political Discourse Analysis: A Method for Advanced Students*. Routledge.
- Falk, S., & van Wynsberghe, A. (2023). Challenging AI for Sustainability: What ought it mean? *AI and Ethics*. <https://doi.org/10.1007/s43681-023-00323-3>

- Fisher, M. (2012). What Is Hauntology? *Film Quarterly*, 66(1), 16–24. <https://doi.org/10.1525/fq.2012.66.1.16>
- Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.
- Fourcade, M., & Gordon, J. (2020). Learning Like a State: Statecraft in the Digital Age. *Journal of Law and Political Economy*, 1(1). <https://doi.org/10.5070/LP61150258>
- Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198237907.001.0001>
- Fung, B. (2023a, May 16). *Mr. ChatGPT goes to Washington: OpenAI CEO Sam Altman testifies before Congress on AI risks* | *CNN Business*. CNN. <https://www.cnn.com/2023/05/16/tech/sam-altman-openai-congress/index.html>
- Fung, B. (2023b, May 17). *How the CEO behind ChatGPT won over Congress* | *CNN Business*. CNN. <https://www.cnn.com/2023/05/17/tech/sam-altman-congress/index.html>
- Gansky, B., & McDonald, S. (2022). CounterFaceTual: How FAccT Undermines Its Organizing Principles. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1982–1992. <https://doi.org/10.1145/3531146.3533241>
- Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. *IEEE Internet Computing*, 21(6), 58–62. <https://doi.org/10.1109/MIC.2017.4180835>
- Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*. <https://doi.org/10.5210/fm.v29i4.13636>
- Goldenfein, J., & Mann, M. (2023). Tech money in civil society: Whose interests do digital rights organisations represent? *Cultural Studies*, 37(1), 88–122. <https://doi.org/10.1080/09502386.2022.2042582>
- Gong, S., Liu, X., & Jain, A. K. (2020). *Mitigating Face Recognition Bias via Group Adaptive Classifier* (No. arXiv:2006.07576). arXiv. <https://doi.org/10.48550/arXiv.2006.07576>
- Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., & Strait, A. (2024). *Auditing Work: Exploring the New York City algorithmic bias audit regime* (No. arXiv:2402.08101). arXiv. <https://doi.org/10.48550/arXiv.2402.08101>
- Hajer, M. A. (2009). *Authoritative Governance: Policy Making in the Age of Mediatization*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199281671.001.0001>
- Hansen, S. S. (2021). Public AI imaginaries: How the debate on artificial intelligence was covered in Danish newspapers and magazines 1956–2021. *Nordicom Review*, 43(1), 56–78. <https://doi.org/10.2478/nor-2022-0004>
- Hawes, B., & Hall, D. W. (2023, December). *After the Summit: Progress in public policy on AI* [Monograph]. University of Southampton. <https://doi.org/10.5258/SOTON/WSI-WP007>
- Hendrycks, D., & Mazeika, M. (2022). *X-Risk Analysis for AI Research* (No. arXiv:2206.05862). arXiv. <https://doi.org/10.48550/arXiv.2206.05862>
- Hess, D. J. (2014). When Green Became Blue: Epistemic Rift and the Corraling of Climate Science. In *Fields of Knowledge: Science, Politics and Publics in the Neoliberal Age* (Vol. 27, pp. 123–153). Emerald Group Publishing Limited. <https://doi.org/10.1108/S0198-871920140000027012>
- Hickman, L., & Serlin, D. (2018). Towards a crip methodology for critical disability studies. In *Interdisciplinary Approaches to Disability*. Routledge.
- Hogan, M. (2020). The Data Center Industrial Complex (2021). *Saturation: An Elemental Politics* (Duke University Press). https://www.academia.edu/39043972/The_Data_Center_Industrial_Complex_2021
- Irani, L. (2015). The cultural work of microwork. *New Media & Society*, 17(5), 720–739. <https://doi.org/10.1177/1461444813511926>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), Article 9. <https://doi.org/10.1038/s42256-019-0088-2>
- Jones, M., & McKelvey, F. (2024). Deconstructing public participation in the governance of facial recognition technologies in Canada. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-024-01952-w>
- JUST-AI. (2023). *Prototyping Ethical Futures Series: 2021 Convening on Futures*. JUST AI. <https://www.just-ai.net/copy-of-2021-convening-futures>
- Katz, Y. (2020). Artificial Whiteness: Politics and Ideology in Artificial Intelligence. In *Artificial Whiteness*. Columbia University Press. <https://doi.org/10.7312/katz19490>
- Kearnes, M., & Wynne, B. (2007). On Nanotechnology and Ambivalence: The Politics of Enthusiasm. *NanoEthics*, 1(2), 131–142. <https://doi.org/10.1007/s11569-007-0014-7>
- Knight, S., Shibani, A., & Vincent, N. (2024). Ethical AI governance: Mapping a research ecosystem. *AI and Ethics*. <https://doi.org/10.1007/s43681-023-00416-z>
- Lasswell, H. D. (1971). *A pre-view of policy sciences*. American Elsevier Publishing Company. <http://archive.org/details/previewofpolicys0000lass>
- Lea, T. (2020). *Wild policy: Indigeneity and the unruly logics of intervention*. Stanford University Press.
- Li, F., Kakhki, A. M., Choffnes, D., Gill, P., & Mislove, A. (2016, 16). Classifiers Unclassified: An Efficient Approach to Revealing IP Traffic Classification Rules. *Proceedings of the 16th ACM Internet Measurement Conference*. Internet Measurement Conference, Santa Monica, CA. <http://david.choffnes.com/pubs/ClassifiersUnclassified-IMC16.pdf>
- Lipsky, M. (2010). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service, 30th Anniversary Expanded Edition* (30th Anniversary, Expanded ed. edition). Russell Sage Foundation.

- Luka, M. E., Harvey, A., Hogan, M., Shepherd, T., & Zeffiro, A. (2015). Scholarship as Cultural Production in the Neoliberal University: Working Within and Against 'Deliverables.' *Studies in Social Justice*, 9(2), Article 2. <https://doi.org/10.26522/ssj.v9i2.1138>
- Marres, N. (2005). *Issues Spark a Public into Being: A Key But Often Forgotten Point of the Lippmann-Dewey Debate* (B. Latour & P. Weibel, Eds.; pp. 208–217). MIT Press.
- McKelvey, F. (2018). *Internet daemons: Digital communications possessed*. University of Minnesota Press.
- McKelvey, F., & Roberge, J. (2023). Recursive Power. In S. Lindgren (Ed.), *Handbook of critical studies of artificial intelligence* (pp. 21–32). Edward Elgar Publishing.
- Menn, J., & Nix, N. (2023, December 7). Big Tech funds the very people who are supposed to hold it accountable. *Washington Post*. <https://www.washingtonpost.com/technology/2023/12/06/academic-research-meta-google-university-influence/>
- Metz, R. (2021, March 24). *Google offered a professor \$60,000, but he turned it down. Here's why*. CNN. <https://www.cnn.com/2021/03/24/tech/google-ai-ethics-reputation/index.html>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2). <https://doi.org/10.1177/2053951716679679>
- Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: Barriers, enablers and next steps. *AI & SOCIETY*, 38(1), 411–423. <https://doi.org/10.1007/s00146-021-01308-8>
- Nagy, P., & Neff, G. (2024). Conjuring algorithms: Understanding the tech industry as stage magicians. *New Media & Society*, 26(9), 4938–4954. <https://doi.org/10.1177/14614448241251789>
- Naiseh, M., Bentley, C., Ramchurn, S., Williams, E., Awad, E., & Alix, C. (2022). Methods, Tools and Techniques for Trustworthy Autonomous Systems (TAS) Design and Development. *Companion of the 2022 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 66–69. <https://doi.org/10.1145/3531706.3536459>
- Newman-Griffis, D., Rauchberg, J. S., Alharbi, R., Hickman, L., & Hochheiser, H. (2023). Definition drives design: Disability models and mechanisms of bias in AI technologies. *First Monday*. <https://doi.org/10.5210/fm.v28i1.12903>
- Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., & Guston, D. (2013). A Framework for Responsible Innovation. In *Responsible Innovation* (pp. 27–50). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118551424.ch2>
- Palm s, K., & Surber, N. (2022). Legitimation crisis in contemporary technoscientific capitalism. *Journal of Cultural Economy*, 15(3), 373–379. <https://doi.org/10.1080/17530350.2022.2065331>
- Paul, R. (2022). Can critical policy studies outsmart AI? Research agenda on artificial intelligence technologies and public policy. *Critical Policy Studies*, 16(4), 497–509. <https://doi.org/10.1080/19460171.2022.2123018>
- Readings, B. (1997). *The University in Ruins*. Harvard University Press.
- S tra, H. S., Coeckelbergh, M., & Danaher, J. (2022). The AI ethicist's dilemma: Fighting Big Tech by supporting Big Tech. *AI and Ethics*, 2(1), 15–27. <https://doi.org/10.1007/s43681-021-00123-7>
- Schneiders, E., Chamberlain, A., Fischer, J. E., Benford, S., Castle-Green, S., Ngo, V., Kucukyilmaz, A., Barnard, P., Row Farr, J., Adams, M., Tandavanitj, N., Devlin, K., Mancini, C., & Mills, D. (2023). TAS for Cats: An Artist-led Exploration of Trustworthy Autonomous Systems for Companion Animals. *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, 1–5. <https://doi.org/10.1145/3597512.3597517>
- Sch n, D. A., & Rein, M. (1994). *Frame reflection: Toward the resolution of intractable policy controversies*. New York : BasicBooks. <http://archive.org/details/framereflectiont00dona>
- Simpson, J. (2024, May 4). Capacity crunch: Why the UK doesn't have the power to solve the housing crisis. *The Observer*. <https://www.theguardian.com/business/article/2024/may/04/capacity-crunch-why-the-uk-doesnt-have-the-power-to-solve-housing-crisis>
- Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation Is not a Design Fix for Machine Learning. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–6. <https://doi.org/10.1145/3551624.3555285>
- Sloane, M., Moss, E., & Chowdhury, R. (2022). A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, 3(2), 100425. <https://doi.org/10.1016/j.patter.2021.100425>
- Stark, L., & Hoffmann, A. L. (2019). Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. *Journal of Cultural Analytics*. <https://doi.org/10.22148/16.036>
- Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science*, 48(1), 25–56. <https://doi.org/10.1177/0306312717741687>
- Stilgoe, J. (2019). *Who's Driving Innovation?: New Technologies and the Collaborative State*. Springer International Publishing.
- Stone, D. (1997). *Policy paradox: The art of political decision making* (Revised Edition). New York : W.W. Norton & Co. <http://archive.org/details/policyparadoxart0000ston>
- Taylor, L., & Dencik, L. (2020). Constructing Commercial Data Ethics. *Technology and Regulation*, 2020, 1–10. <https://doi.org/10.26116/techreg.2020.001>
- Tharoor, I. (2023, July 26). Analysis | The world reckons with a new 'Oppenheimer moment.' *Washington Post*. <https://www.washingtonpost.com/world/2023/07/26/oppenheimer-nolan-ai-artificial-intelligence/>

- Tuck, E., & Ree, C. (2013). A Glossary of Haunting. In S. H. Jones, T. E. Adams, & C. Ellis (Eds.), *Handbook of autoethnography* (pp. 639–657). Left Coast Press.
- Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & SOCIETY*, 35(1), 147–163. <https://doi.org/10.1007/s00146-018-0845-5>
- Turnbull, N. (2008). Harold lasswell’s “problem orientation” for the policy sciences. *Critical Policy Studies*, 2(1), 72–91. <https://doi.org/10.1080/19460171.2008.9518532>
- Turner, R. H. (2001). Role Theory. In J. H. Turner (Ed.), *Handbook of Sociological Theory* (pp. 233–254). Springer US. https://doi.org/10.1007/0-387-36274-6_12
- UK Government. (2019, June 10). *A guide to using artificial intelligence in the public sector*. GOV.UK. <https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector>
- UK Government. (2024, February 6). *A pro-innovation approach to AI regulation: Government response*. GOV.UK. <https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response>
- van Hulst, M., & Yanow, D. (2016). From Policy “Frames” to “Framing”: Theorizing a More Dynamic, Political Approach. *The American Review of Public Administration*, 46(1), 92–112. <https://doi.org/10.1177/0275074014533142>
- Vicente, P. N., & Dias-Trindade, S. (2021). Reframing sociotechnical imaginaries: The case of the Fourth Industrial Revolution. *Public Understanding of Science (Bristol, England)*, 30(6), 708–723. <https://doi.org/10.1177/09636625211013513>
- Whittaker, M. (2021). The steep cost of capture. *Interactions*, 28(6), 50–55. <https://doi.org/10.1145/3488666>
- Widder, D. G. (2024). Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1295–1304. <https://doi.org/10.1145/3630106.3658973>
- Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12(4), 505–523. <https://doi.org/10.1111/regg.12158>

Strategic misrecognition and speculative rituals in generative AI

Sun-ha Hong

Stanford University, USA

✉ residual@stanford.edu

Abstract

Public conversation around generative AI is saturated with the ‘realness question’: is the software really intelligent? At what point could we say it is thinking? I argue that attempts to define and measure those thresholds misses the fire for the smoke. The primary societal impact of realness question comes not from the constantly deferred sentient machine of the future, but its present form as *rituals of misrecognition*. Persistent confusion between plausible textual output and internal cognitive processes, or the use of mystifying language like ‘learning’ and ‘hallucination’, configure public expectations around what kinds of politics and ethics of genAI are reasonable or plausible. I adapt the notion of *abductive agency*, originally developed by the anthropologist Alfred Gell, to explain how such misrecognition strategically defines the terms of the AI conversation.

I further argue that such *strategic misrecognition* is not new or accidental, but a central tradition in the social history of computing and artificial intelligence. This tradition runs through the originary deception of the Turing Test, famously never intended as a rigorous test of artificial intelligence, to the present array of drama and public spectacle in the form of competitions, demonstrations and product launches. The primary impact of this tradition is not to progressively clarify the nature of machine intelligence, but to constantly redefine values like intelligence in order to legitimise and mythologise our newest machines – and their increasingly wealthy and powerful owners.

Keywords: generative AI; machine intelligence; agency; ritual; spectacle; history of AI

1.1 The real artificial intelligence

Public discourse around generative AI (genAI) is saturated with what we might call the realness question: is the software ‘really’ intelligent, creative, or even, sentient? Does its textual, image, and video output count as ‘really’ artistic or meaningful? The quasi-public releases of Stable Diffusion, Midjourney and DALL-E 2 during the summer of 2022, and then ChatGPT’s sudden launch that November, provoked a cacophony of public spectacles, often explicitly presented as tests of realness. A Midjourney-generated image won an art prize at a local fair (Roose, 2022); GPT-4 allegedly passed bar exams with flying colours (Weiss, 2023); and a United Nations event staged a ‘press conference’ by humanoid robots (AP News, 2023). Yet the realness question does not present any kind of coherent philosophical inquiry; there are rarely clear and consistent criteria for distinguishing genuine intelligence or real creativity from its

facsimiles. Rather, the realness question provides a social ritual for dramatising and renegotiating the boundaries between ourselves and machines – a ritual that can be traced, some cultural histories of automata suggest, back to Western Antiquity (Kang, 2011). In this role, the realness question mediates other dominant public debates around genAI's capabilities and consequences, such as machine intelligence, labour replacement, and 'hallucination'.

These rituals for understanding and evaluating artificial intelligence play out through public dramas. From entrepreneurs' bombastic, self-serving predictions about imminent AI breakthroughs, to enthusiastic users seeking to test genAI tools for signs of intelligence, such spectacles help shape what becomes widely understood as possible, plausible, and inevitable about new technologies. Distinct from the substantive question of whether a given machine is 'really intelligent', public dramas elicit visceral, affective relationships with novel technologies, moulding the sentimental conditions upon which specific debates around regulation or adoption take place. This tendency extends a long historical pattern of how science and technology become publicly understood and justified. Akin to the demonstrative, even theatrical role of experiments in early modern science (Shapin & Schaeffer, 1985) to modern America's fascination with technological 'sublime' (Nye, 1994), public tests of genAI's 'realness' cultivate a collective sense of what proofs or breakthroughs might look and feel like.

Crucially, this dramatisation of genAI repeatedly reinforces a fundamental misrecognition of the technology and its relationship to human qualities like intelligence and creativity. Drawing on Peircian semiotics, Weatherby and Justie (2022) argue that the very question of whether AI 'resembles' human intelligence is a 'metaphysical misstep' based on a 'naïve iconic interpretation' – where visceral resemblance at the output level, such as in LLM-generated news articles or poetry, is taken as proof that machine learning systems process information as human brains do. This misrecognition is not simply an epistemic error, attributable to technological complexity and thus resolvable through literacy or explainability. Rather, it is a strategic misrecognition that calibrates collective expectations around how AI is to be evaluated, and what kinds of political and ethical questions are relevant or plausible – a process that is frequently led by and for a narrow elite of entrepreneurs, venture capitalists, and celebrity scientists. Generative AI output and its various misadventures thus populate public debates with anecdotal folklore, media framing, and other tropes toward the broader myth that genAI is on the cusp of an imminent and inevitable transformation of society (e.g. Bender, 2022; Sadowski, 2018; Vinsel, 2021). Below, I make two arguments: first, that public dramas around genAI's realness question are a central medium for perpetuating this strategic misrecognition; second, that such misrecognition establishes a common set of tests, thresholds, and indicators for evaluating AI and its promises – providing a tautological cycle of justification for the tech elites.

This approach helps clarify the strategic function of the speculative hype and panic around successive generations of technological futures, including generative AI. Much of the current debate around AI's social impact has been structured around the apparent opposition between 'accelerationist' and 'doomerist' positions, and the ostensible feud between the two. Thus Marc Andreessen (2023), one of Silicon Valley's most influential venture capitalists, insists that AI will "save the world" through a virtuous cycle of productivity growth, trickle-down economics and 'skyrocketing' consumer spending. Meanwhile, the 'doomers' issue high-profile calls to pause AI development, such as the longtermist Future of Life Institute's March 2023 letter, featuring star signatories like the computer scientist Yoshua Bengio and Elon Musk (see Gebru et al., 2023). Yet both positions effectively share a foundational commitment to the realness question. Both camps share the foundational assumption that existing genAI systems provide a direct technological ramp to 'really' intelligent machines, i.e. artificial general intelligence [AGI]. It is upon this speculative common ground that both 'camps' argue AGI is imminent and inevitable, and that the incumbent elite of technical AI researchers and industry leaders must continue building these tools at their discretion.

The persistent misrecognition of genAI as 'sparks of AGI' thus supplies pseudo-metaphysical justification for further concentration of power and wealth around AI. These narratives help consolidate

policy debates and investment funding around speculative logics of superintelligence and far future risks (McKelvey & Roberge, 2023, p. 21). Thus Geoffrey Hinton, whose scientific contribution to backpropagation has been central to contemporary machine learning breakthroughs, has since rebranded himself as a public speculator around the existential dangers of AI (e.g. Metz, 2023). He has argued that concerns about AI's amplifying racial disparities or environmental destruction "aren't as existentially serious" and that we should rather focus on the robot apocalypse (Bender, 2023). This dynamic extends what Meredith Whittaker (2021) has called the capture of AI ethics by industry actors – in which Silicon Valley in particular has used its political and economic capital to direct resources and visibility towards incremental, techno-solutionist renditions of AI ethics, squeezing out more foundational problems around facial recognition and algorithmic decision-making systems. The notion of strategic misrecognition emphasises how accelerationist and doomer predictions, and other public dramas around the realness question, work to "[re]capture power and create obfuscation by making new myths and legends" (Coldicutt, 2023) around speculative future technologies and their incumbent elites.

1.2 Dramas of justification

Public dramas around generative AI echo a far longer pattern of metaphysical and teleological myths in the history of technology. In the early United States, technology was quickly enlisted toward national origin stories of 'second creation' through settler mastery of nature (Nye, 2003). Post-colonial nation-building often mobilises grand technological projects – dams, telecom infrastructure, and so on – into belief-building objects that Paula Chakravartty (2004) calls 'temples of the future'. Silicon Valley's decades-long fascination with posthumanism and the Singularity extends this mythological tapestry and especially its eschatological dimension (Dinerstein, 2006; Hong, 2022; Kneese, 2015; F. Turner, 2016). Speculative discourse around human extinction and radical transcendence, across technologies like AGI and space exploration (Roberts & Hogan, 2019; Smith & Burrows, 2021; Utrata, 2023), carve out bounded spaces for rehearsing the social and political implications of technology.

These are what Bryan Pfaffenberger (1992) has called technological dramas: public events and controversies through which technological change is collectively visualised, alternative interpretations confronted, providing a scaffolding of expectations and attitudes around speculative notions like artificial general intelligence. Technological dramas are strongly ritualistic in the sense that these collective, staged performances work to enact new social realities (e.g. Rappaport, 1999; V. Turner, 1976). It is the formalised roles and scripts in a ritual, and its public and staged quality, that produces a collective sense that the change in question (say, the reclassification of a neophyte into a full member of the community) is being appropriately tested and subsequently resolved. Today, the means of production and propagation of these dramas is largely dominated by the incumbent group of tech megacorporations and investors, from Microsoft through their relationship with OpenAI to figures like Andreessen.

Public dramas around genAI reinforce the primacy of the realness question as an interpretive lens and historical narrative. The design and promotion of tools like ChatGPT and Bard has encouraged the general public to experiment, 'testing' the machines for the appearance of intelligence or artistry. Casey Fiesler notes that ChatGPT's breakout success owed much to its anthromorphising interface – a deliberate design choice and not driven by any technical necessity – which has the cost of obfuscating the models, data, and human labour behind the front-end (in Bender & Fiesler, 2023). Investment and product development decisions at the corporate level are also driven by the need to generate and steer public dramas in order to bootstrap underdeveloped technologies into cultural and market prominence. David Kampmann (2024) has shown how Silicon Valley's reliance on venture capital funding produces hard economic motivations for such dramatisation: since investment is rooted in expectations of massive 'blitzscaling' success for a few breakout products, spectacles of supposedly imminent breakthroughs become a key collateral through which to gain further funding and contracts. OpenAI's development of ChatGPT in late 2022 was itself frantically rushed (Roose, 2023) specifically to pre-empt competitors in shaping public expectations

around genAI. This speed was rewarded with extended media and policy attention, which OpenAI's leaders maximised with a months-long world tour meeting state leaders, culminating in an early 2024 share sale valuing the company at 80 billion USD. Public dramas, in other words, are key to how nascent and unreliable technologies become embedded into cultural and financial futures.

The realness question thus employs dramatic means (around the misrecognition of 'real' machine intelligence) towards strategic ends (of shaping AI's regimes of justification). Below, I analyse genres of drama around genAI as intelligent machines, and situate them in a longer history of computing and AI – eventually arriving in the long, zombified afterlife of the Turing Test. These dramas around genAI do not constitute a collective, open-ended exploration for the meaning of new technologies, but an elite-driven pursuit of cultural justification for already determined strategies of market domination. In *The New Spirit of Capitalism*, Luc Boltanski and Eve Chiapello (2007) explain that one of capitalism's distinct strengths is at crafting its own regimes of justification: the publicly accepted 'tests' and standards by which new paradigms like neoliberal privatisation or AI-driven automation might be considered proven. For instance, there is mounting evidence that the use of predictive algorithms for human behaviour and social decision-making is beset by fundamental flaws: the target of prediction, such as criminal risk, cannot be robustly defined and operationalised, and the very conceit of 'predicting' social outcomes lacks a coherent theoretical foundation (see overview in Wang et al., 2023). Yet the use of single-figure accuracy scores, for example, has effectively positioned these tools as engaged in a race with the 'average human', yielding a contest in which victory provides an objective seal of legitimacy. This is one example of the justificatory regime built through the incessant testing of generative AI for 'real' intelligence: a collective framework against which other theories and values must make themselves legible – or become relegated to the margins.

This dramatic approach allows us to focus on the concrete societal consequences of genAI, which are driven not by a metaphysical ascent up the ladder of general intelligence, but by the ambiguous, arbitrary, and underdefined ways in which we are encouraged to talk and think about the technology. Researchers have painstakingly shown how present-day generative AI does not meaningfully understand language (Bender, 2023); are stochastic parrots that recombine text and image without meaning (Gebu et al., 2023; Jiang et al., 2023), and exhibits no direct pathway to developing 'general' intelligence (Marcus, 2022) – despite the very label of 'artificial intelligence' imbuing contemporary machine learning techniques with 'magical thinking' (Raley & Rhee, 2023, p. 188) around its capabilities and historical trajectory.¹ This definitional messiness is a feature, not a bug, of the drama. Tech industry and its elites have proven consistently skilled at deflecting scrutiny from regulators, academia and civil society, often by encouraging debates and tests that are predisposed towards incremental, technical solutions: is there a 'tradeoff' between security and privacy in digital systems? Could a more 'optimised' algorithm avoid racial bias and other historical inequalities? (e.g. see critiques in Goldenfein et al., 2020; McKelvey & Neves, 2021; Turow et al., 2015) Accounts of strategic misrecognition can help us think more deliberately and independently about the regimes of justification around generative AI.

Below, I focus on the misrecognition of intelligence in genAI discourse. The field of artificial intelligence has famously lacked a clear and consensual definition of intelligence for its entire history. In public dramas around genAI and the future of AI technology, intelligence appears both as a master criterion for progress and as a messy, inconsistent notion overlapping with qualities like consciousness and creativity. Public dramas around genAI rely on such overlaps: debates around specific capabilities (such as writing generic news articles) or other humanlike qualities (such as performing empathy) quickly become scenes for pseudo-metaphysical speculation around machine intelligence. This allows actors to strategically gesture towards wider cosmological speculations, seeking to endow today's clunky machines with an aura of historical agency. I return below to the history of artificial intelligence as a

¹ Invoking philosopher of science Helen Verran, Lucy Suchman calls it a 'hardening of the categories' – where the fixity of the sign exudes a semiotic confidence belied by the instability of the classificatory work now concealed by the sign (Suchman, 2023, p. 3).

strategically underdefined term. But first: what exactly is being ‘misrecognised’ when genAI is subject to tests of ‘real’ machine intelligence?

2.1 The abduction of generative AI

In 1997, the British anthropologist Alfred Gell received a terminal cancer diagnosis – and duly resolved to complete as quickly as possible his final argument: that art objects exhibit agency through a process of *abduction* (Gell, 1998). In this view, many kinds of art – a painted portrait, a finely carved Trobriand canoe, apotropaic emblems like good-luck charms – may all exhibit agency insofar as they are *attributed* causal power in their cultural context. Drawing on Peircian semiotics, Gell argued that art objects are indexes, “chiefly function[ing] as a sign that points to something else” (Chua & Elliott, 2013, p. 8). The objects do not possess agency, but are imputed agentic capacity by others. Thus agency is a relational *effect*, and not an essential *quality*; it is something that “rub[s] off on objects” rather than emanating from within (Abram, 2005, p. 13).²

Agency, of course, is not intelligence. But genAI’s misrecognition of intelligence involves a very similar form of abductive attribution. Software *output* is consistently misrecognised as *process*: the appearance of *bildungsroman* form or impressionist technique in the output is taken as proof of some form of underlying skill, intention, or understanding – despite growing evidence of basic methodological flaws behind many high-profile demonstrations that sought to prove LLMs can ‘know’ and ‘think’ (Martínez, 2024; also see Heaven, 2023b).³ Notably, Gell emphasises that abductive agency does not require free will or consciousness, only a social object with a basic capacity to mediate causal effects. Decoupling agency from metaphysics in this way unclutters much of the category confusion in the realness question. Since agency is abducted, i.e. socially attributed as relational effect rather than an internally possessed quality, we can recognise genAI output’s obvious impact on human beings – to bewilder, to convince, to affect – without needing to jump to logically incoherent claims about machine intelligence. We also need not presume that abductive activity is always a sign of user ignorance or superstition, if we realise that abduction is a common and ordinary form of human sensemaking.

In 2022, Google engineer Blake Lemoine went public with claims that, based on his experience with the company’s LaMDA chatbot, the chatbot must be sentient. The extended global media attention on the drama – heightened by Google’s subsequent termination of Lemoine – further reinforced the dominance of the realness question in public debates around generative AI (e.g. see Johnson, 2022). Here, output is again misrecognised as process, and relational phenomena used to speculate about essential qualities. The Lemoine saga extends a long pattern of such abductive panics in the history of computing. Chatbots, in particular, have been dominated by attributions of humanlike intelligence and consciousness for decades. In 1966, the public debut of the ELIZA chatbot famously provoked similar misrecognition around machine intelligence. Joseph Weizenbaum, its creator, suggested that his own secretary imputed human qualities onto the bot – despite the immediately visible limitations of its Sokratic tendency to repeat the user’s questions back at them (Weizenbaum, 1976, p. 6). The irony was that Weizenbaum had designed ELIZA in part to show how relatively unintelligent software with a small vocabulary could still simulate the appearance of conversational interaction.

Such misrecognitions are often presented as a true/false test: is the computer really intelligent (that is, possessing it as an essential and innate quality), or is the user simply mistaken? Abduction provides an alternative interpretation: that such attributions are pragmatic responses that help establish *useful* fixtures for social and moral responses to new technology, even if they may be technically incorrect. Blaming

² More broadly, relational approaches have been central to accounts of information processing and decision-making in complex human-machine assemblages, from drone warfare to biohacking and smart urbanisms, where human perception, cognition and judgment are thoroughly interwoven with the machines around them (Hayles, 2017; Sampson, 2017).

³ Consider OpenAI’s widely publicised claim that GPT-4 scored in the 90th percentile in the Uniform Bar Exam. An independent evaluation by Martínez (2024) has shown that the percentile figure was estimated using unspecified methods, which appear to draw from a skewed dataset of human scores.

algorithms for bias, for example, can be a (temporarily) effective way to seek some accountability for a harmful result where traditional attributions of blame have been obfuscated by new decision-making systems. For tech elites, misrecognition has been a profitable way to stoke controversy around the possible sentience of AI systems, persuading investors, policymakers and the public that AI development must be a grand civilisational priority above all other concerns.

Gell had insisted that a proper anthropology of art needed to be strictly divorced from the question of aesthetics. There were two reasons for such a move. First, the causal effects attributed to art objects were not always dependent on aesthetic appreciation. Second, this would help us consider abduction in objects that don't easily fit into modern Western norms around art – the primacy of painting, the hierarchy of artistic value via museums and curatorial control, the norms of contemplative appreciation, and so on. Similarly, setting aside aesthetic considerations helps clarify what is happening in an abductive process: genAI output's ability to fool or amaze human observers is not any necessary indicator of intelligence, creativity or consciousness, but only of its capacity for inviting abduction. Such capacity emerges from, above all, genAI systems' ability to ingest massive amounts of human art and communication – and the webs of referentiality, intention, and meaning invested into them – and to break down and reprocess it all into an industrial swill for mass production (Salvaggio, 2023). It is from this cacophony of fragments that we infer meaning, relating a meaningless image or text back into our learned sense of art history of literary genre. As scholars have already noted of 'big data' systems, when there is enough noise, it becomes easy enough to 'discover' signal apophenically (e.g. Steyerl, 2016). This distinction puts a stopper on the slippery slope of reasoning, where it is assumed *if* genAI were to produce 'real' art (or legal opinion, or scientific paper), then the machine would need to be treated as an *artist* – and the ensuing speculations about machines' ability to replace human labour wholesale or their eligibility for various human rights. As Gell (1988, chapter 7) pointed out, many objects across world cultures are abductively attributed with powerful causal effects without conferring virtuosic status on their creators.

Thus, if a genAI-produced poem should unsettle us or bring us to tears, this says nothing much about essential humanlike qualities, but points to a long tradition of abductively artistic objects that have made humans laugh and cry without necessarily moving the bar on sentience or existential risk. Automata, in particular, have historically incited public engagement specifically in the form of 'detective work', provoking speculations of realness: does this machine reason? Does it simulate something of life? And if so, what does it suggest about our own reasoning and living? (Geoghegan, 2020) One emblematic case is Jacques de Vaucanson's 18th century invention, *Le Canard Digérateur* [The Digesting Duck]. Vaucanson claimed that the machine had mastered the ability to eat, and subsequently, shit – and showed off the routine in popular demonstrations in and around 18th century Paris. All available evidence suggests that the duck had simply stored pre-made shit pellets inside; but what is telling is that well into the 19th century, commentators often assumed that it must have possessed artificial anatomical features, such as a mechanical intestine (Riskin, 2003). Machines have been objects of abduction for a long time, and misrecognition of this abductive relation has long fuelled speculation around what makes *us* intelligent.

2.2 Uncanny capture

Why, though, does genAI output prove so fertile for abductive attribution? Let us remain a little longer with Gell, who provides the label of the causal milieu (1998, p. 37): a cultural, interpretive context that shapes what kinds of agency is likely to be anticipated and recognised. From sorcery and cursework to carnivalesque effigies, the milieu regulates cultural norms around how some forms of abductive attribution become acceptable, while others become categorised as superstitious or transgressive. Gell himself cites the case of the *Rokeby Venus* in 1914 (Gell, 1998, Chapter 4), in which suffragette Mary Richardson's slashing a venerable Velázquez is condemned as a barbaric attack rather than being seen as an 'artistic' experiment of defacement. The difference between what constitutes 'genuine' acts of creativity and skill, and what constitutes mere copy or even vandalism, depends on this quasi-

Bourdieuian milieu. In their very design as well as the promotional discourses around them, genAI tools like Midjourney seek to shape a milieu for themselves, offering statistical recombination of image data as the new benchmark of intelligence.

Experiences of mystery, awe, and captivation are central to shaping these causal milieus. Here we return to technology as myth, as temples of the future. Gell suggests that abduction of agency often occurs when we perceive ordinary means achieving extraordinary results. The brushwork on a Vermeer or the carvings on a Trobriand canoe is not so alien as to defy *any* interpretation, but sufficiently extraordinary to evoke a sense of mystery as well – especially if we’ve been exposed to modern art museums and other causal milieus in our lives (Gell, 1998, p. 71). GenAI output, especially when designed to respond to iterative and experimental human prompts, incites similar activities of speculation and wonder through the aforementioned ‘detective work’. Such abductive activity is further encouraged by the tech elite-driven narrative that AI is fundamentally ‘unexplainable’ even to its own human creators – an air of mystique that is then used to justify both utopian and dystopian claims about the technology’s imminently superintelligent status. Microsoft researchers’ much-publicised paper, ‘Sparks of AGI’ (Bubeck et al., 2023), thus argued that GPT-4 exhibits some forms of ‘general intelligence’, even while employing definitions and measures of intelligence that they acknowledged were rather arbitrary and subjective – and, as other researchers quickly pointed out, relied on an infamous 1997 editorial that was written as an activist endorsement of innate racial differences in human intelligence (Bender, 2023).

The mystery around genAI output thus provokes a sense of the *uncanny*, furnishing space for speculative meaning-making and abductive attribution. In computing, the uncanny usually means the uncanny valley: the zone in which synthetic images are realistic enough to invite recognition *as* real, but not quite realistic enough to convince. Yet both in Masahiro Mori’s original 1970 formulation and contemporary use, the uncanny valley functions as a useful label for clearly observable phenomena, but there is no robust theory of what makes computer graphics uncanny under what conditions. More broadly, the uncanny has a rich tradition in philosophy and literary theory, and has often been central to cultural responses to automation throughout history. The surrealists turned to cyborgs and other uncanny intersections of humans and machines to process the ‘shock of industrialisation’ (Foster, 1991) – as the situationists would later do in response to the dominant myth of cybernetic futures (Routhier, 2023). Many moral panics around new media technologies can also be understood as a reaction to the perceived unsettling of boundaries and categories (also see Ravetto-Biagioli, 2019; Kang, 2011). We may think of the ambiguous gap between my ‘offline’ and ‘online’ self, and later, my ‘data double’ or ‘data doppelganger’ (e.g. Cheney-Lippold, 2017; Watson, 2014), or between bots and ‘real’ users – and the discomfort that arises from the eventual recognition that there can often be no firm separation between the two.

The uncanny is thus an important site for inciting wonder and terror around genAI objects, and in turn, practices of abductive attribution. Susan Lepselter (2016) writes that the uncanny is a central affect in many everyday myths about politics and society. It is a percolating sense, never quite concretised into official forms, but recurring in the margins, in the patterns of the stories we tell: for instance, “a sense that life in America [today] is shaped by some ineffable, enormous power, a power that can be seen only in the patterns of its effects” (p. 1) – from relatively discredited powers like aliens and the Illuminati, to relatively normalised ones like ‘the economy’ or ‘the (deep) state’. The key is that the uncanny is not simply an experience of the destruction of the scaffolding of meaning, but also part of a constant, everyday effort to rebuild that scaffolding. Drawing on Freud’s *unheimlich* – usually read as ‘unhomely’ or ‘unfamiliar’, but also, in his own writing, involving forms of ‘involuntary repetition’ – Lepselter (pp. 22–23) argues that there is a habitual element to the uncanny: that we scramble to *unearth* (also see Vidler, 1999, p. 27) some uncanny gap, and then to see it again and again, so that we may once again seek to close or ‘complete’ it. This provocative, restless dimension comes to the fore in the abductive speculations around generative AI.

In September 2022, multimedia artist Supercomposite reported ‘discovering’ a memorably creepy, deformed face while experimenting with negative prompt weights in genAI tools (in which the software seeks to generate output that is *opposite* to the description) (e.g. Rose, 2022). It was an emergent effect: since this particular face was not specifically summoned by a positive prompt or name, the human user could only indirectly control its appearance in generated output. But reappear it did across many different generated images, with a consistently macabre stylisation. (It is apropos that the face often takes a doll-like aesthetic: Freud examined dolls as a key example of *unheimlich*, drawing from Ernst Jentsch’s earlier reading of the doll Olympia in Romantic writer E.T.A. Hoffmann’s 1817 story, *Der Sandmann*.) Supercomposite promptly named the face ‘Loab’ – observing that even when Loab-producing prompts were combined with prompts for entirely different shapes or styles, human viewers could easily recognise an enduring ‘Loabness’ in the resulting output. The mystery around Loab’s provenance, and its memorable visual features, provoked familiar speculations and panic around genAI’s intelligence, memory, and even spirituality. Once again, the realness question had provided a shared blueprint for collective practices of speculation around machine intelligence.

3.1 The spectacle of the test

How are causal milieus formed and maintained for generative AI, cultivating a wider public ‘literacy’ on how to read the technology for ostensible signs of intelligence? One key medium is the *test* as spectacle. From art competitions to bar exams, the realness question (and its uncanny affects) is often operationalised into public tests, presenting putatively objective reference points for tracking the progress of artificial intelligence. Predictions of the imminent arrival of AGI thus faithfully recycle an old tradition of spectacles, tests, and predictions. In February 2022, Sam Altman opined that “techno-optimism is the only good solution to our current problems” – and that this supposedly ‘radical’ and unpopular position will deliver ‘limitless, cheap, clean energy’, as well as space colonisation, fusion, and ‘cure all human disease’ (Sam Altman [@sama], 2022). He has continued to leverage public attention around OpenAI’s generative tools to spread the teleological rhetoric of abundance. Such speculations are not lackadaisical asides, but a central part of the emerging regime of justification around AI. The tech elite’s more ‘pessimistic’ public declarations have tended to exactly mirror Altman’s rhetoric, with only the outcomes inverted: a Center for AI Safety statement in June 2023 argued that “mitigating the risk of extinction from AI should be a global priority” precisely because its superhuman prowess and universal adoption is virtually guaranteed (*Statement on AI Risk*, 2023; also see Wong, 2023). It is no contradiction, in that sense, that Altman was one of the signatories for the letter: AI optimism and pessimism are joined at the hip of strategic interest, protecting a tightly bounded sphere of technological futures.

Such predictions liberally remix religious tropes out of context to craft eschatological messages of urgency and inevitability (also see Cheney-Lippold, 2024; Gebru & Torres, 2024). Thus Anthony Levandowski, one of the most prominent engineers in the self-driving car research community, is also the founder of the short-lived “Way of the Future,” a self-proclaimed “church of AI” – founded in 2015 and shuttered in 2020 – having sought to prepare humanity for the imminent arrival of superintelligent machines. Ray Kurzweil, a serial prophet, has now spent several decades predicting the Singularity as the impending next great chapter of civilisation – a vision of total technological rewriting of society. Implicit in these tales is the conceit that only the Valley’s unique community of transgressive geniuses can see into this future and prepare the rest of us for its arrival (also see Hong, 2022).

Predicting some form of ‘genuine’, all-purpose artificial intelligence in the near future, quietly stepping over the expiry date, and then predicting it again, has been a regular ritual throughout the history of AI research. In 1970, Marvin Minsky, widely hailed as a founding figure of AI, told *Life* magazine that a machine will possess ‘general intelligence of an average human being’ in ‘three to eight years’ (Darrach, 1970) – despite the field’s perennial lack of a robust, consensual definition of intelligence throughout its entire existence (e.g. Dick, 2019). In many ways, “the field [of AI] has always involved fantasy in search

of a practical method” – a conjectural bet, made for promotional reasons as well as intellectual, that machine intelligence might be possible (Stark, 2023a, p. 367). Minsky and his colleagues specifically chose the term artificial intelligence to cast the broadest possible net for funding and significance. Yarden Katz describes how John McCarthy, in proposing the 1956 Dartmouth workshop today mythologised as the founding event for AI, eventually settled on the term in order to avoid “embracing MIT mathematician Norbert Wiener, who coined [cybernetics], as the [new] field’s ‘guru’.” (Katz, 2020, p. 22) Although Claude Shannon had initially argued against the undefined and loaded use of ‘intelligence’, he eventually signed on, along with Minsky, as the workshop’s co-organiser. The new moniker quickly achieved its aims: ARPA enthusiastically funded artificial intelligence projects, sidelining cyberneticians despite their complaints of the AI researchers’ ‘con trick’ (Katz, 2020, pp. 24–25).

This pattern of speculation and prediction is not restricted to a discrete ‘marketing’ dimension, hermetically sealed away from the more serious work of actually doing science and technology. Communities of AI and computing have for decades organised their collective sense of technological progress and its direction through genres of public tests and spectacle. One obvious correlate here are performance competitions, such as ImageNet’s Large Scale Visual Recognition Challenge (ILSVRC) or the Netflix Prize, which organise a formal conflict through which specific models may emerge victorious, and more importantly, a common sense of technical progress may be coordinated. Such tests extend a wider technoscientific culture of justification, from benchmarks (Raji et al., 2021) to prototypes (Corsín Jiménez, 2014; F. Turner, 2016) and variants of design thinking (Irani, 2018). Similarly, a ‘test-bed’ approach is central to the smart technology paradigm: populations and cities become sites of constant, real-time adjustments and experiments in which models are always already rolling out changes into the environment and thus requiring indefinite adaptation (Halpern & Mitchell, 2022). In short, the very notion of artificial intelligence emerged in part as a bid to shape the public conditions of justification. The regular production of speculative predictions about transformative futures, and more formalised rituals of tests, challenges and prototypes, both serve as what Latour (1988) called the ‘theatre of the proof’, cultivating common expectations around what can realistically be demanded of technology in the first place.

3.2 Turing’s mirage

One of AI’s most popular origin myths, of course, is a test. In 1950, Alan Turing proposed ‘the imitation game’. Its details are well known: X converses with Y and Z, physically separated and without somatic markers like voice. If X cannot distinguish which of Y and Z is human or machine, the machine has passed the test. This fairly specific exercise was then glossed with a rather speculative estimate: that ‘in about fifty years’ time’, computers would have sufficient storage capacity to pass the test (Turing, 1950).

Crucially, Turing did *not* address the question, ‘can machines think?’, which he explicitly dismissed as “too meaningless to deserve discussion” (p. 442). The imitation game was instead about *substituting* the problem of intelligence with a pragmatic and provisional task. In the 1950 paper, Turing entertained some familiar objections – for example, that machines cannot produce novel thought, or fall in love – but the aim was not to defeat these objections, only to suggest that these objections themselves are not conclusive enough to dismiss the possibility of machine intelligence in the future (pp. 442–454). Simply put, the Turing Test was never a *test*; in Turing’s own words, it remained a game, a demonstration (also see Mitchell, 2024).

The central role of *imitation* in the Turing Test sabotages any retroactive effort to formalise the test as a watertight proof of intelligence, as many have discovered over the decades. After all, imitation points to computers as fundamentally abductive agents, rather than possessing any essentialised agency. Yet the imitative dimension was not accidental; it was rooted in Turing’s lifetime of work. Well before the 1950 paper, the Turing Machine had proposed a universal mathematical definition of computability, and its iconic tape mechanism was designed to mimic human mathematical cognition (Hodges, 2009). One might speculate an alternative history of AI where the field seriously engaged the relational implications of

imitation in the Turing Test – leading, perhaps, to more realistic theories of machine interactivity unburdened of the metaphysical hot air around the sentience of text and pixel recombination software.

Computing took the other road. Since 1950, the Turing Test has taken on a mythological life across both the lay public and expert communities, often taken as an ur-test, a final frontier which offers the possibility of absolutely proving machine intelligence. Although many researchers across philosophy, computer science and other domains have dissected the test's limitations, it has remained prominent both as a thought experiment and as a concrete research objective. Some competitions, such as the Loebner Prize (1990-2020), took a fairly literal path, each year inviting participants in a standardised Turing Test. Some observers were enthusiastic, others more skeptical (e.g. Floridi et al., 2009; Marcus, 2014) – while, somewhat ironically given his own historical promotion of utopian predictions about AI, Marvin Minsky loudly decried the Prize as a publicity stunt. James Moor, a prominent moral philosopher in computing and AI who wrote extensively on the Test for over four decades, admitted that the test fails to provide an operational definition of AI – while also insisting that the test can somehow provide a philosophical foundation for the question of machine intelligence (Moor, 2001).

This is not to paint a history in which original nuance is diluted over time. Simone Natale (2021) argues that since its inception, the Turing Test helped establish deception – or, in my terms here, misrecognition – as a constitutive element of artificial intelligence. Turing had readily admitted that the 1950 piece was written as a “form of propaganda that would stimulate philosophers, mathematicians, and scientists to engage more seriously with the machine intelligence question.” (Natale, 2021, pp. 19–20) This ‘deceptive’ strategy determined the test's initial design as well as its subsequent legacy. Consider the actual parameters of the test. Turing's design requires an ‘average interrogator’ without special knowledge (of, say, linguistics or anthropology), without additional contextual knowledge of how machines of this kind tend to be designed, and tasked with guessing the machine after ‘five minutes of questioning’. There is little theoretical justification as to why any of these parameters are necessarily or ideal. Instead, Turing chose to present a kind of test which can provide an effective and immediate theatre of proof.⁴

The legacy of the Turing Test, too, belongs primarily in the annals of technoscientific demonstrations towards constructing public regimes of justification. Researchers have suggested that competitions like the Netflix Prize taken on a sport-like dimension in terms of its coverage and cultural impact (e.g. Orr & Kang, 2023); we may also point to more literal sport events staged for AI justification, such as AlphaGo's duel against Go master Lee Sedol in 2016. As computer scientist Delip Rao (2023) has argued, these benchmarks and competitions also exert a gravitational pull for research and funding, creating perverse incentives in what kinds of ‘breakthroughs’ are pursued and celebrated.

There has been no shortage of candidates for dethroning the Turing Test with other tests. But most efforts play with the parameters while retaining its dramatic structure. Some suggest a Lovelace Test, for instance. Taking up Ada Lovelace's comment on Babbage's Analytical Engine that computers cannot ‘originate’, it is argued that a test for more creative content, such as poetry, would provide a satisfactory test of intelligence (e.g. Bringsjord et al., 2003). Yet the misrecognition of an abductive relation as an essential one remains. The search for the ‘right’ or better test thus extends rather than breaks from this wider culture of ritualised, spectacular demonstrations. Leif Weatherby (2023) relays Lydia Liu's observation that “all ‘Turing-style’ tests reduce intelligence to a competition between humans and machines” – a sport in which the winner takes home the prized status of public proof and recognition.

What might a qualitatively different test do for public conversations around generative AI? Jack Stilgoe (2023) argues for a Weizenbaum Test: if a technology *can* be built, *should* it? Who will benefit, and who will bear the costs? Such a test steps outside the myopic laboratory conditions imposed by the Turing Test and its heirs, in which an individual human specimen is locked inside a (Chinese) room for experimental proof. Outside this strange and unnecessary confinement, it becomes clear that where

⁴ Turing recommended that when given an arithmetic question in an imitation game, the computer should “pause about 30 seconds” before giving the answer (Turing, 1950, p. 434) – a parlour trick quite comparable to how present-day genAI systems take to folksy turns of phrase to perform humanlikeness.

computers are attributed by (some) human beings with qualities like intelligence or consciousness, that practice of attribution is itself shaped by arcs of political and economic interest that have cleared the social stage for the arrival of the new technology in the first place (Marvin, 1988; Pfaffenberger, 1992).

4. Conclusion: Devaluation machines

The misrecognition surrounding generative AI is ritually reproduced abduction, rather than incidental forms of human ignorance. It serves strategic objectives for shaping the terms of the AI conversation in favour of the incumbent tech elites, rather than simply a stumbling block on the road to sentient machines. GenAI, in other words, should be evaluated primarily as an abduction machine with a social and political strategy. Tools like Midjourney and ChatGPT are deliberately designed to produce a particular form of abduction, derived from a long history of anthromorphisation and of a fascination with uncanny machines.

The most serious consequence of such abduction – and one in urgent need of critical analysis and reframing today – is that these machines enact a widespread *devaluation* of the human qualities that the technology purports to reproduce. The question of whether a machine can ‘replace’ a human at a task is a misleading one, because it reproduces the realness question as a substantive problem, and implies that the answer depends on how ‘smart’ the machine is. Rather, the abduction view asks how machines dramatise and redefine qualities like intelligence – and in doing so, justify the pernicious suggestion that whatever can cheaply and fleetingly deceive as intelligence, or creativity, or empathy, must be just as good as the real thing.

Consider the use of genAI to undermine the social valuation of creative and artistic labour. Luke Stark turns to 20th century animation as a parallel shift in this regard. Animation, too, was a technological transformation of image production strongly predicated on the abductive function. “All animation simplifies, and so is implicitly dependent in the human ability to make meaningful heuristic inference” (2023b). In this view, animation and generative AI become effective by *provoking* abductive attribution. Yet this is only half of the process. As the animated output becomes misrecognised as agentic, they “pus[h] the the actual living labour of their animators into the background or offstage” (ibid.). Optimising genAI’s abductive capacity involves relentlessly calculating *whatever output might provoke desired user inference at the lowest cost*. This optimisation occurs to a degree in the training phase, where models are honed to select whatever combination of textual and image data that ‘looks real enough’ to the low-paid, subcontracted human coder (e.g. Dzieza, 2023). But it also occurs in the public presentation of recombinatory bullshit as ‘real’ art or communication. The corollary to the argument that genAI output could replace human workers is the premise that nothing deeper than the rough appearance of humanlike qualities has any value.

This politics of devaluation thus corresponds to a rather retrograde essentialism that I described as a strategic misrecognition about abduction: the belief that rather than being complex and relational things, intelligence or creativity is an essence that an AI might ‘have’ or not have. This misrecognition is not simply an epistemic confusion, but a crucial instrument for the ongoing politics of justification around data-driven technologies as mechanisms of extraction (e.g. Thatcher et al., 2016). Generative AI as a market product grows by attaching itself to existing infrastructures of production and then offering a lower-cost alternative. From illustration and digital journalism to essay mills and bot farms, genAI takes a pre-existing milieu for assigning value to labour, and by decoupling output from process, promises to discard the bulk of that labour and its associated costs. Where Amazon Mechanical Turk and successive crowdwork platforms once achieved this arbitrage largely by slicing employment into smaller chunks of time in order to minimise wages and benefits (Irani, 2015; Shapiro, 2018), generative AI offers a sort of Turing Test for the precariat: that if a generated text or image is *just* good enough to pass as human-made for a non-discerning customer in the short term, then the human skill and labour for that work must not have much value at all.

The public dramas around genAI thus shape the likely path and extent of the technology's political economic impacts. Anticipating and critically evaluating those 'disruptions' require tracing how dramatised myths till the soil of public and industry expectation. What researchers have called fauxtimation or 'potemkin AI' (e.g. Sadowski, 2018; Taylor, 2018) is not a temporary stepping stone, but the most likely endpoint of an optimisation function that prioritises minimally convincing output above meaning, intention, or process. As genAI relies on abductive effects to justify and promote itself, it is likely to incentivise other economic and social decisions that strip away such 'extraneous' processes. In this sense, an image generator software does not 'replace' artists or designers; rather, it executes a scorched earth policy of devaluation, depressing the social value of that work for all parties.

In 1979, Lyotard opened *The Postmodern Condition* by noting that in computerised societies, we may thus expect a thorough exteriorisation of knowledge with respect to the "knower," at whatever point he or she may occupy in the knowledge process. The old principle that the acquisition of knowledge is indissociable from the training (*Bildung*) of minds, or even of individuals, is becoming obsolete and will become ever more so (Lyotard, 1984, p. 5; also see Beardon, 1994).

The postwar rise of computing technologies corresponded with a broader shift from reason to rationality: at the broadest strokes, a move away from liberal subjects endowed with relatively self-sufficient capabilities for reasoning and knowledge, toward impersonal and calculative systems as custodians of knowledge (Erickson et al., 2013). The very notion of an 'artificial' intelligence extends this epistemic project, and its premise that intelligence can be decoupled from human subjects. Yet generative AI and its public dramas are also deeply reliant on the enduring belief that in order to think, to know, to write, one must be a subject. This vestigial imagination is now being exploited to advance the strategic misrecognition of machine intelligence, where the appearance of thinking, knowing, and writing, provokes speculation that these machines should perhaps be treated as subjects. For Kittler, the phonograph had already inaugurated the age of "writing without a subject" (1986, p. 44); today, the spectre of the writing subject still haunts the fantasy of the autonomous machine.

References

- Abram, S. (2005). Introduction: Science/technology as politics by other means. *Focaal—European Journal of Anthropology*, 45, 3–20.
- Andreessen, M. (2023, June 6). *Why AI Will Save the World*. Andreessen Horowitz. <https://a16z.com/ai-will-save-the-world/>
- Beardon, C. (1994). Computers, postmodernism and the culture of the artificial. *AI & Society*, 8(1), 1–16. <https://doi.org/10.1007/BF02065174>
- Bender, E. (2022). *On NYT Magazine on AI: Resist the Urge to be Impressed*. Medium. <https://medium.com/@emilymenonbender/on-nyt-magazine-on-ai-resist-the-urge-to-be-impressed-3d92fd9a0edd>
- Bender, E. M. (2023, July 6). Talking about a 'schism' is ahistorical. *Medium*. <https://medium.com/@emilymenonbender/talking-about-a-schism-is-ahistorical-3c454a77220f>
- Bender, E. M., & Fiesler, C. (2023, March 1). *The Limitations of ChatGPT with Emily M. Bender and Casey Fiesler*. The Radical AI Podcast. <https://www.radicalai.org/chatgpt-limitations>
- Boltanski, L., & Chiapello, E. (2007). *The New Spirit of Capitalism*. Verso.
- Bringsjord, S., Bello, P., & Ferrucci, D. (2003). Creativity, the Turing Test, and the (Better) Lovelace Test. In J. H. Moor (Ed.), *The Turing Test: The Elusive Standard of Artificial Intelligence* (pp. 215–239). Springer Netherlands. https://doi.org/10.1007/978-94-010-0105-2_12
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. <https://doi.org/10.48550/arXiv.2303.12712>
- Chakravartty, P. (2004). Telecom, National Development and the Indian State: A Postcolonial Critique. *Media, Culture & Society*, 26(2), 227–249. <https://doi.org/10.1177/016344370404117>
- Cheney-Lippold, J. (2017). *We are Data—Algorithms and the making of our digital selves*. New York University Press.
- Cheney-Lippold, J. (2024). The silicon future. *New Media & Society*, OnlineFirst. <https://doi.org/10.1177/14614448241234864>
- Chua, L., & Elliott, M. (Eds.). (2013). *Distributed objects: Meaning and mattering after Alfred Gell*. Berghahn.

- Coldicutt, R. (2023, June 1). *On Understanding Power and Technology*. Medium. <https://rachelcoldicutt.medium.com/on-understanding-power-and-technology-1345dc57a1a>
- Corsín Jiménez, A. (2014). Introduction: The prototype: More than many and less than one. *Journal of Cultural Economy*, 7(4), 381–398. <https://doi.org/10.1080/17530350.2013.858059>
- Dick, S. (2019). Artificial Intelligence. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.92fe150c>
- Dinerstein, J. (2006). Technology and Its Discontents: On the Verge of the Posthuman. *American Quarterly*, 58(3), 569–595. <http://dx.doi.org/10.1353/aq.2006.0056>
- Dzieza, J. (2023, June 20). *Inside the AI Factory*. The Verge. <https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots>
- Erickson, P., Klein, J. L., Daston, L., Lemov, R., Sturm, T., & Gordin, M. D. (2013). *How Reason Almost Lost Its Mind*. University of Chicago Press.
- Floridi, L., Taddeo, M., & Turilli, M. (2009). Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest. *Minds and Machines*, 19(1), 145–150. <https://doi.org/10.1007/s11023-008-9130-6>
- Foster, H. (1991). Exquisite Corpses. *Visual Anthropology Review*, 7(1), 51–61. <https://doi.org/10.1525/var.1991.7.1.51>
- Gebru, T., Bender, E. M., McMillan-Major, A., & Mitchell, M. (2023, March 31). *Statement from the listed authors of Stochastic Parrots on the "AI pause" letter*. Distributed AI Research Institute. <https://www.dair-institute.org/blog/letter-statement-March2023/>
- Gebru, T., & Torres, É. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, 29(4). <https://doi.org/10.5210/fm.v29i4.13636>
- Gell, A. (1998). *Art and Agency: An Anthropological Theory*. Clarendon Press.
- Geoghegan, B. D. (2020). Orientalism and Informatics: Alterity from the Chess-Playing Turk to Amazon's Mechanical Turk. *Ex-Position*, 43, 45.
- Goldenfein, J., Green, B., & Viljoen, S. (2020, April 17). *Privacy Versus Health Is a False Trade-Off*. Jacobin. <https://jacobinmag.com/2020/4/privacy-health-surveillance-coronavirus-pandemic-technology>
- Halpern, O., & Mitchell, R. (2022). *The smartness mandate*. MIT Press.
- Hayles, K. N. (2017). *Unthought: The Power of the Cognitive Nonconscious*. University of Chicago Press.
- Heaven, W. D. (2023, August 30). *AI hype is built on high test scores. Those tests are flawed*. MIT Technology Review. <https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/>
- Hodges, A. (2009). Alan Turing and the Turing Test. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (pp. 13–22). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_2
- Hong, S. (2022). Predictions Without Futures. *History and Theory*, 61(3), 371–390. <https://doi.org/10.1111/hith.12232>
- Irani, L. (2015). The cultural work of microwork. *New Media & Society*, 17(5), 720–739. <https://doi.org/10.1177/1461444813511926>
- Irani, L. (2018). "Design Thinking": Defending Silicon Valley at the Apex of Global Labor Hierarchies. *Catalyst: Feminism, Theory, Technoscience*, 4(1), Article 1. <https://doi.org/10.28968/cftt.v4i1.29638>
- Jiang, H. H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., & Gebru, T. (2023). AI Art and its Impact on Artists. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374. <http://dx.doi.org/10.1145/3600211.3604681>
- Johnson, K. (2022, June 14). LaMDA and the Sentient AI Trap. *Wired*. <https://www.wired.com/story/lamda-sentient-ai-bias-google-blake-lemoine/>
- Kampmann, D. (2024). Venture capital, the fetish of artificial intelligence, and the contradictions of making intangible assets. *Economy and Society*, 0(0), 1–28. <https://doi.org/10.1080/03085147.2023.2294602>
- Kang, M. (2011). *Sublime Dreams of Living Machines: The Automaton in the European Imagination*. Harvard University Press.
- Katz, Y. (2020). *Artificial whiteness: Politics and ideology in artificial intelligence*. Columbia University Press.
- Kittler, F. A. (1986). *Gramophone, Film, Typewriter* (G. Winthrop-Young & M. Wutz, Trans.). Stanford University Press.
- Kneese, T. (2015, October). Spiritual Machines: Transhumanist Temporalities and Startup Culture. *Humanism and Its Prefixes*.
- Latour, B. (1988). The Pasteurization of France. In *Practical Synthetic Organic Chemistry: Reactions, Principles, and Techniques*. Harvard University Press.
- Lepselter, S. (2016). *The Resonance of Unseen Things: Poetics, Power, Captivity, and UFOs in the American Uncanny*. University of Michigan Press.
- Liotard, J.-F. (1984). *The Postmodern Condition: A Report on Knowledge*. University of Minnesota Press.
- Marcus, G. (2014, June 9). What Comes After the Turing Test? *The New Yorker*. <https://www.newyorker.com/tech/annals-of-technology/what-comes-after-the-turing-test>
- Marcus, G. (2022, December 25). What to Expect When You're Expecting ... GPT-4 [Substack]. *Marcus on AI*. <https://garymarcus.substack.com/p/what-to-expect-when-youre-expecting>

- Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*.
<https://doi.org/10.1007/s10506-024-09396-9>
- Marvin, C. (1988). *When Old Technologies Were New: Thinking About Electric Communication in the Late Nineteenth Century*. Oxford University Press.
- McKelvey, F., & Neves, J. (2021). Introduction: Optimization and its discontents. *Review of Communication*, 21(2), 95–112.
<https://doi.org/10.1080/15358593.2021.1936143>
- McKelvey, F., & Roberge, J. (2023). Recursive power: AI governmentality and technofutures. In *Handbook of Critical Studies of Artificial Intelligence* (pp. 21–32). Edward Elgar Publishing.
- Metz, C. (2023, May 1). 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead. *The New York Times*.
<https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html>
- Mitchell, M. (2024). The Turing Test and our shifting conceptions of intelligence. *Science*, 385, eaqd9356.
- Moor, J. H. (2001). The Status and Future of the Turing Test. *Minds and Machines*, 11(1), 77–93.
<https://doi.org/10.1023/A:1011218925467>
- Natale, S. (2021). *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test*. Oxford University Press.
- Nye, D. (2003). *America as Second Creation*. MIT Press.
- Nye, D. (1994). *American Technological Sublime*. MIT Press.
- Orr, W., & Kang, E. (2023, November 9). *AI as a sport: On the competitive epistemologies of benchmarking*. Society for the Social Studies of Science [4S] Annual Meeting, Honolulu.
- Pfaffenberger, B. (1992). Technological Dramas. *Science, Technology & Human Values*, 17(3), 282–312.
<https://doi.org/10.1177/016224399201700302>
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the Everything in the Whole Wide World Benchmark. *NeurIPS 2021*.
- Raley, R., & Rhee, J. (2023). Critical AI: A Field in Formation. *American Literature*, 95(2), 185–204.
<https://doi.org/10.1215/00029831-10575021>
- Rao, D. (2023, December 20). The Great AI Weirding [Substack newsletter]. *AI Research & Strategy*.
<https://deliprao.substack.com/p/the-great-ai-weirding>
- Rappaport, R. (1999). *Ritual and Religion in the Making of Humanity*. Cambridge University Press.
- Ravetto-Biagioli, K. (2019). *Digital Uncanny*. Oxford University Press.
- Riskin, J. (2003). The Defecating Duck, or, the Ambiguous Origins of Artificial Life. *Critical Inquiry*, 29(4), 599–633.
<https://doi.org/10.1086/377722>
- Roberts, S. T., & Hogan, M. (2019). Left Behind: Futurist Fetishes, Prepping and the Abandonment of Earth. *B2o: An Online Journal*, 4(2).
- Roose, K. (2022, September 2). An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. *The New York Times*.
<https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html>
- Roose, K. (2023, February 3). How ChatGPT Kicked Off an A.I. Arms Race. *The New York Times*.
<https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html>
- Rose, J. (2022, September 7). Why Does This Horrifying Woman Keep Appearing in AI-Generated Images? *Vice*.
<https://www.vice.com/en/article/g5vjw3/why-does-this-horrifying-woman-keep-appearing-in-ai-generated-images>
- Routhier, D. (2023). *With and Against: The Situationist International in the Age of Automation*. Verso.
- Sadowski, J. (2018, August 6). *Potemkin AI*. Real Life. <http://reallifemag.com/potemkin-ai/>
- Salvaggio, E. (2023, October 29). *The Hypothetical Image*. Cybernetic Forests.
<https://www.cyberneticforests.com/news/social-diffusion-amp-the-seance-of-the-digital-archive>
- Sam Altman [@sama]. (2022, February 4). *Techno-optimism is the only good solution to our current problems. Unfortunately, somehow expressing optimism about the future has become a radical act.* [Tweet]. Twitter.
<https://twitter.com/sama/status/1489648210842816516>
- Sampson, T. D. (2017). *The Assemblage Brain: Sense Making in Neuroculture*. University of Minnesota Press.
- Shapin, S., & Schaeffer, S. (1985). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press.
- Shapiro, A. (2018). Between autonomy and control: Strategies of arbitrage in the “on-demand” economy. *New Media & Society*, 20(8), 2954–2971. <https://doi.org/10.1177/1461444817738236>
- Smith, H., & Burrows, R. (2021). Software, Sovereignty and the Post-Neoliberal Politics of Exit. *Theory, Culture & Society*, 38(6), 143–166. <https://doi.org/10.1177/0263276421999439>
- Stark, L. (2023a). Breaking Up (with) AI Ethics. *American Literature*, 95(2), 365–379. <https://doi.org/10.1215/00029831-10575148>
- Stark, L. (2023b, March 16). *ChatGPT is Mickey Mouse*. Daily Nous. <https://dailynous.com/2023/03/16/philosophers-on-next-generation-large-language-models/>
- Statement on AI Risk. (2023, June 2). Center for AI Safety. <https://www.safe.ai/statement-on-ai-risk#signatories>
- Stilgoe, J. (2023). We need a Weizenbaum test for AI. *Science*, 381(6658), eadk0176. <https://doi.org/10.1126/science.adk0176>

- Suchman, L. (2023). The uncontroversial ‘thingness’ of AI. *Big Data & Society*, 10(2).
<https://doi.org/10.1177/20539517231206794>
- Steyerl, H. (2016). A sea of data: apophenia and pattern recognition. *e-flux*, 72. <http://www.e-flux.com/journal/a-sea-of-data-apophenia-and-pattern-mis-recognition/>
- Taylor, A. (2018). The Automation Charade. *Logic*, 5. <https://logicmag.io/failure/the-automation-charade/>
- Thatcher, J., O’Sullivan, D., & Mahmoudi, D. (2016). Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space*, 34(6), 990–1006.
<https://doi.org/10.1177/026377581663319>
- Turing, A. (1950). Computing Machinery and Intelligence. *Mind, New Series*, 59(236), 433–460.
<https://doi.org/10.1093/mind/LIX.236.433>
- Turner, F. (2016). Prototype. In B. Peters (Ed.), *Digital Keywords: A Vocabulary of Information, Society & Culture* (pp. 256–268). Princeton University Press.
- Turner, V. (1976). Social Dramas and Ritual Metaphors. In R. Schechner & M. Schumann (Eds.), *Ritual, Play, and Performance: Readings in The Social Sciences / Theatre* (pp. 97–120). The Seabury Press.
- Turow, J., Hennessy, M., & Draper, N. (2015). *The Tradeoff Fallacy: How Marketers Are Misrepresenting American Consumers And Opening Them Up To Exploitation*. Annenberg School for Communication, University of Pennsylvania.
- UN tech agency rolls out human-looking robots for questions at a Geneva news conference. (2023, July 7). AP News.
<https://apnews.com/article/humanoid-robots-better-leaders-ai-geneva-486bb2bad260454a28aaa51ea31580a6>
- Utrata, A. (2023). Engineering Territory: Space and Colonies in Silicon Valley. *American Political Science Review*, 1–13.
<https://doi.org/10.1017/S0003055423001156>
- Vidler, A. (1999). *The Architectural Uncanny: Essays in the Modern Unhomely*. MIT Press.
- Vinsel, L. (2021). *You’re Doing It Wrong: Notes on Criticism and Technology Hype*. Medium. <https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5>
- Wang, A., Kapoor, S., Barocas, S., & Narayanan, A. (2023). Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 626. <https://doi.org/10.1145/3593013.3594030>
- Watson, S. M. (2014, June 16). *Data Doppelgängers and the Uncanny Valley of Personalization*. The Atlantic.
<http://www.theatlantic.com/technology/archive/2014/06/data-doppelgangers-and-the-uncanny-valley-of-personalization/372780/>
- Weatherby, L. (2023, November 13). *ChatGPT broke the Turing test*. The Boston Globe.
<https://www.bostonglobe.com/2023/11/13/opinion/turing-test-ai-chatgpt/>
- Weatherby, L., & Justie, B. (2022). Indexical AI. *Critical Inquiry*, 48(2), 381–415. <https://doi.org/10.1086/717312>
- Weiss, D. C. (2023, March 16). *Latest version of ChatGPT aces bar exam with score nearing 90th percentile*. ABA Journal.
<https://www.abajournal.com/web/article/latest-version-of-chatgpt-aces-the-bar-exam-with-score-in-90th-percentile>
- Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W.H. Freeman and Company.
- Whittaker, M. (2021). The Steep Cost of Capture. *ACM Interactions*, November-December, 51–55.
<https://doi.org/10.1145/3488666>
- Wong, M. (2023, June 2). *AI Doomerism Is a Decoy*. The Atlantic.
<https://www.theatlantic.com/technology/archive/2023/06/ai-regulation-sam-altman-bill-gates/674278/>

Getting democracy wrong

How lessons from biotechnology can illuminate limits of the Asilomar AI principles

Gwendolyn Blue¹, Mél Hogan²

¹ University of Calgary, Calgary, Alberta, Canada

² Queen's University, Kingston, Ontario, Canada

✉ ggblue@ucalgary.ca

Abstract

Recent developments in large language models and computer automated systems more generally (colloquially called ‘artificial intelligence’) have given rise to concerns about potential social risks of AI. Of the numerous industry-driven principles put forth over the past decade to address these concerns, the Future of Life Institute’s Asilomar AI principles are particularly noteworthy given the large number of wealthy and powerful signatories. This paper highlights the need for critical examination of the Asilomar AI Principles. The Asilomar model, first developed for biotechnology, is frequently cited as a successful policy approach for promoting expert consensus and containing public controversy. Situating Asilomar AI principles in the context of a broader history of Asilomar approaches illuminates the limitations of scientific and industry self-regulation. The Asilomar AI process shapes AI’s publicity in three interconnected ways: as an agenda-setting manoeuvre to promote longtermist beliefs; as an approach to policy making that restricts public engagement; and as a mechanism to enhance industry control of AI governance.

Keywords: Principles, Asilomar, Governance, Artificial Intelligence, Biotechnology, Longtermism

1. Introduction

In 2017, the Future of Life Institute (FLI) unveiled the Asilomar AI principles aimed at guiding AI research, development, and commercialization toward the benefit of all (Stirling, 2018). These principles were crafted at an invitation-only workshop at the Asilomar Conference Grounds in Pacific Grove, California which brought together prominent figures in industry and academia with a shared objective to create a unified vision for AI governance.

Founded in 2014 by tech entrepreneurs and academic researchers,¹ the FLI supports and promotes longtermism, an influential belief system, social movement, and policy narrative that focuses on ensuring

¹ FLI founders include Skype co-founder Jaan Tallin, DeepMind’s Viktoriy Krakovna, MIT cosmologist Max Tegmark, and physicist Anthony Aguirre at University of California Santa Cruz. Advisors include entrepreneurs such as Elon Musk, AI researchers such as Stuart Russell, physicists such as Steven Hawking (now deceased), geneticists such as George Church, and science communicators such as Alan Alda, among others (Labaree, 2014).

the well-being of inhabitants of the distant future by preventing catastrophic risks of technological innovation in the present (Torres 2021, Crary 2023). Drawing inspiration from philosopher Nick Bostrom's concept of effective altruism, a moral belief that advocates for wealth creation as the best way to promote the public good (Bostrom 2014), William MacAskill coined the term longtermism to describe approaches that seek to protect future humans from existential risks that threaten humanity's long-term potential, including risks from AI (MacAskill, 2022). The FLI fulfils its longtermism mission, in part, by organizing invite-only conferences with AI experts to promote principles to guide the future of AI. As stated on the FLI website, "we believe that meetings of experts from industry and academia, and from across the globe, can significantly improve common understanding and improve our collective future" (FLI, n.d.).

The FLI's Asilomar AI principles are one illustration of broader AI ethics initiatives, which are principle-based approaches to governance led by industry, government, academia, and not-for-profit groups that have gained momentum over the past decade (Jobin et al., 2019; Marchant et al., 2020; Milmo, 2023; Munn, 2023). Examples of AI ethics statements include civil society-driven initiatives such as Amnesty International's Toronto Declaration, state-driven initiatives such as the European High Level Expert Group on AI's Ethics Guidelines for Trustworthy AI, and stakeholder initiatives such as the Montréal Declaration on Responsible AI, among many others (for comprehensive overview, see Jobin et al., 2019). Ethical guidelines are a response to concerns about the social implications of emerging technologies, and are often presented as a flexible alternative to, or prototype for, government regulation to guide the trajectory of technological research, development, and implementation in ways that do not stymie innovation. Critics argue that principle-based ethical initiatives are a strategic manoeuvre on the part of technology developers and private industry to sidestep government regulation and avoid public scrutiny (Wagner, 2018; Bender et al., 2021; McKelvey and Roberge, 2023; Munn, 2023). The inherent ambiguity of principles further compounds the challenge of their implementation.

The Asilomar AI principles were among the first public-facing principles to guide the development of AI and played an agenda-setting role for subsequent principle-based frameworks (Fjeld et al., 2020). To date, the principles have been adopted and signed by thousands of individuals and organizations. This paper positions the Asilomar AI principles as a noteworthy dimension of AI's contemporary publicity. Situating the Asilomar AI principles within a longer history of development of the Asilomar model, a well-established principle-based approach to science policy, highlights the FLI's emphasis on self-regulation and particularly the assumption that the people who develop technology and put it to use are best positioned to define the terms of public debate and government regulation. Our concern is not with principle-based governance of emerging technologies per se, but rather with use of the Asilomar model by the FLI to promote its guiding principles. First developed in the early 1970s in response to innovations in genetic engineering, the Asilomar model – as will be discussed further – has since been applied to other publicly contentious technologies such as geoengineering, synthetic biology, human gene editing, and most recently, AI. While proponents laud the Asilomar model as a successful example of expert solidarity in anticipating social consequences of technology, critics highlight problematic implications including positioning publics and government regulators in a reactive and reactionary position, forever playing catch-up with technological advances (Hurlbut, 2015a; Hurlbut, 2015b; Jasanoff and Hurlbut, 2018; Taylor and Dewsbury, 2019).

What lessons can be learned about AI's current publicity by situating the Asilomar AI principles in a longer history of Asilomar-style approaches to scientific and technological governance? The FLI's initiatives share many similarities with Asilomar approach to technology governance, including an invitation-only conference which resulted in the release of guiding principles, followed by calls for a temporary pause on AI research to provide time for experts to develop guidelines and regulation. A closer look reveals that, in focus and intent, the Asilomar AI principles, and the process to develop them, co-opted problematic aspects of the Asilomar model. In what follows, we first trace a brief history of debates about experiments in expert-driven technical regulation dating back to the initial Asilomar conferences

on recombinant DNA which set a precedent for technological elites to anticipate and address potential social consequences of research, and to forward their own research and commercial interests. Next, we examine how the FLI's Asilomar AI process, which included the Beneficial AI conference, the Asilomar AI principles, and a follow-up letter calling for a pause on AI research, drew on the original Asilomar model and, in so doing, shaped AI's publicity in three interconnected ways: as an agenda-setting manoeuvre to promote longtermist beliefs; as an approach to policy making that restricts public engagement; and as a mechanism to enhance industry control of AI governance.

2. Setting the stage: The Asilomar conferences on recombinant DNA

In the early 1970s, scientists were among the first to sound the alarm about the potential social consequences of emerging genetic engineering technologies, concerns which led to Asilomar-style approaches to science policy. Since details about the Asilomar conferences are well-documented in the academic literature, we present here a brief overview of key events.

In 1972, biochemist Paul Berg, along with other colleagues at Stanford University, proposed the creation of a new DNA molecule by inserting genetic material from a monkey virus (SV40) into a virus that infects *Escherichia coli*, a bacteria found in the human gut (Jackson et al., 1972). The new hybrid molecule, and the technique used to develop it, are commonly referred to as recombinant DNA (rDNA).² In January 1973, in response to scientific and public concerns about the potential consequences of rDNA research, Berg organized a scientific conference at the Asilomar Hotel and Conference Grounds in Pacific Grove, California to discuss potential biohazards with a focus on laboratory safety protocols. In the same year, Stanley Cohen (Stanford) and Herbert Boyer (UCSF) published results of the first successful cloning experiment which introduced rDNA into bacteria, for which they would later receive the first biotechnology patent which inaugurated the biotechnology industry. In June 1973, further concerns about rDNA research were raised at the Gordon Conference on Nucleic Acids, resulting in a letter by scientists at the conference to the National Academies of Science (NAS) and the National Academy of Medicine to implement guidelines for biotechnology research (Singer and Soll, 1973). In response, NAS created an ad hoc study group chaired by Berg called the Committee on Recombinant DNA Molecules, which penned an open letter calling for a moratorium on rDNA research until the hazards were better understood (Berg et al., 1974). This letter had two important outcomes: it resulted in the first voluntary moratorium in the history of molecular biological research; and, it raised awareness among scientists and broader publics of potential harmful implications of biotechnology.

The second, and more well-known, Asilomar conference took place at the same location in February 1975. Lasting over four days, this meeting - officially called the International Congress on Recombinant DNA Molecules - convened an international group of 140 scientists primarily in molecular biology, along with a handful of lawyers and journalists, to discuss potential consequences of rDNA, and to develop guidelines to lift the voluntary moratorium. Following the conference, the organizers filed a report to NAS outlining recommendations and conditions under which rDNA research could proceed (Berg et al., 1975). These conditions included the establishment of appropriate safeguards, namely efforts to ensure the containment of newly created organisms through the implementation of biological and physical barriers. In 1976, these containment principles were formulated into National Institute of Health (NIH) guidelines for the funding of rDNA research.

Twenty-five years after the initial Asilomar conferences on rDNA, lawyer and medical ethicist Alexander Capron organized a follow up workshop - Asilomar 3 - to discuss whether Asilomar's legacy had contemporary relevance for emerging technology innovations (Barinaga, 2000). Attended by many

² Recombinant DNA is now a routine and commonly used laboratory technique.

organizers of the original Asilomar conferences, Asilomar 3 convened a more diverse gathering including scientists but also philosophers, historians, ethicists, legal scholars, and government officials (Capron and Shapiro, 2001). The attendees agreed that the original Asilomar model was no longer appropriate for the following reasons. Views about biotechnology had become polarized in ways that made consensus difficult, if not impossible, to achieve. Controversies over genetically modified foods in the 1990s revealed that scientific proclamations of safety-through-containment were insufficient to quell public concerns about commercial, ethical, and social consequences of genetic engineering. In turn, powerful commercial forces in genomic science created conflicts of interest that undermined the legitimacy of scientific self-governance (Barinaga, 2000). Conference attendees concluded that these limitations signalled the need to extend the Asilomar model to include a broader range of perspectives, values, and expertise beyond scientists and technology developers to ensure that social, ethical, and legal aspects of emerging technologies were addressed in advance of implementation (Rufo and Ficorilli, 2019).

Following the Asilomar 3 meeting, several Asilomar + (“self-regulation plus”) models were applied to contentious technologies such as synthetic biology (Ferber, 2004), geo-engineering (ASOC, 2010), environmental algorithms (Galaz, 2015), and human germline editing (Baltimore et al., 2015). At one high-profile Asilomar+ meeting in 2015, three years before He Jiankui used genome editing on human embryos, scientists, bioethicists, and legal experts convened to discuss the technical and social implications of human germline editing made possible by advances in gene editing technologies. Berg was instrumental in organizing this workshop which was modelled in the style of the original Asilomar conference but with attention to the need for a more open dialogue with ethicists about the steps to guarantee the safe development of human genome modification (Baltimore et al., 2015). Concurrently, editorials in academic venues such as *Nature* (Editorial, 2015) and public-facing venues such as the *New York Times* (Capron, 2015) highlighted the continued relevance of the Asilomar model for science policy and underscored the necessity of engaging with broader stakeholders beyond the scientific community.

The Asilomar model continues to play a symbolic role in science and technology policy as shorthand for responsible self-regulation, where scientists and technology developers take the lead in publicly acknowledging potential harmful consequences, and in developing principles to guide technology development and implementation. Asilomar’s continued relevance stems in part from the successful outcomes of the original Asilomar conferences, outcomes which include improved attention to laboratory safety, and, from the perspective of scientists, an avoidance of stringent government regulation (Berg, 2004, 2006, 2008). In Berg’s estimation, the success of the Asilomar model resulted specifically from focused discussion of plausible risks that could be feasibly addressed by technical experts (Berg, 2004).

The Asilomar model has also generated significant critique, particularly about the democratic implications of limiting input to technical experts and restricting the scope of issues to technical matters (Rogers, 1975; Krimsky, 1982; Gottweis, 2005; Gregorowius et al., 2017; Hurlbut, 2015a; Hurlbut, 2015b; Jasanoff et al., 2019; Parthasarathy, 2015; Rufo and Ficorilli, 2019; Taylor and Dewsbury, 2019). In essence, the Asilomar model enables technology developers to powerfully shape the trajectory of regulation and public opinion. The Asilomar model – even the improved Asilomar + approach – does not empower diverse constituencies to shape technological research and innovation agendas, but rather secures support for expert-driven governance and technology commercialization (Rufo and Ficorilli, 2019). According to Ben Hurlbut, the Asilomar model gets democracy wrong: arguably, from a democratic perspective, technologies should be shaped by collectively articulated visions of the public good and the public interest, and not solely by the people who stand to profit from technology development and implementation (Hurlbut, 2015a). Asilomar-style models, including Asilomar+, position technical experts as the most qualified to regulate the social aspects of technology, and in turn position publics and regulatory agencies as passive recipients of principles generated by scientific and technical communities. “The public role that the Asilomar story celebrates is one of dependence,” as Taylor and Dewsbury note, “with the public passively learning — and deferring to — science’s authoritative judgment about what is at stake and when a democratic reaction is warranted” (2019: 12).

Many critical commentators have subsequently concluded that Asilomar and Asilomar+ models are a missed opportunity to anticipate and address social concerns in a democratic fashion (Parthasarathy, 2015). Proposals for improvement include the development of procedures to enhance public engagement in developing ethical principles (Rufo and Ficorilli, 2019), and, more substantively, the creation of new global institutions to bring diverse views to inform the governance of emerging technologies (Jasanoff and Hurlbut, 2018).

3. Asilomar AI: Process and principles

The Asilomar AI Principles were developed during an FLI conference on Beneficial AI that took place in January 2017 over two and a half days at the Asilomar Conference Center, as a sequel to a 2015 conference held in Puerto Rico. The California-based Beneficial AI conference was attended by over 1000 participants, including several leading AI researchers and entrepreneurs, in addition to experts from economics, law, ethics, and philosophy. Many of the speakers at the conference are well-known promoters of longtermism such as Nick Bostrom, Ray Kurzweil, and Elon Musk. The conference schedule, and all talks, are publicly available on the FLI website (FLI, n.d.).

Session descriptions provide an indication of the focus of the Beneficial AI conference, much of which centred on the permissibility of technological innovation, ways to circumvent regulation, and the inevitability of super-intelligent AI. Sessions ranged from economics (“How can we grow our prosperity through automation without leaving people lacking income or purpose”), creating human-level AI (“What remaining obstacles can be identified?”), superintelligence (“What can we do now to maximize the probability of a positive outcome”), law, policy, and ethics (“How can we update legal systems, international treaties and algorithms to be more fair, ethical, and efficient and to keep pace with AI?”).

A key outcome of the Beneficial AI workshop was the development of a consensus document consisting of 23 voluntary principles organized into three categories (Research Issues, Ethics and Values, Long-Term Issues) (FLI, 2017). These principles represent a voluntary commitment on the part of its signatories for the research, development, and commercialization of AI. Overall, the Asilomar AI principles reflect the techno-optimist, longtermist beliefs of the FLI. Consider, for instance, the opening statement of the Asilomar AI principles: “Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles will offer amazing opportunities to help and empower people in decades and *centuries* ahead” (FLI, 2017: *italics added*). Underpinning these principles is the assumption that AI will develop superintelligence, and that AI is inherently beneficial to all humans (if governed correctly).

As others have observed, many of principles are vague and inconsistent, and we present here select examples for illustrative purposes (see Stapf-Fine et al., 2018 for more extensive overview). For instance, the first principle states that the goal of AI research should be to create beneficial but not undirected intelligence. This connection is weak: something can be undirected (i.e. uncontrolled) but still be beneficial, and something can be controlled and still be harmful. Principles three and four call for a culture of cooperation, marked by “constructive and healthy exchanges” among AI researchers and policy-makers. What makes for a “healthy exchange” of views between researcher and policy-makers? Cooperation is not inherently healthy, as conflict and conflicting opinions also play a significant role in policy development. It is questionable whether policy makers should be expected to cooperate and align harmoniously with the interests of industry. Principle 12 requires the right to personal data protection but is restricted to data generated by users and does not address data collected about users. Principle 14 aims to empower and benefit as many people as possible, suggestive of an instrumental goal of maximized self-interest consistent with longtermism, but with no indication of how such benefits might be realized in an inequitable world. Overall, the Asilomar AI principles paint a utopian picture of the inevitability and desirability of AI, suggestive of the possibility of a dystopian future if AI is not developed in alignment with longtermist values. Beyond questions of logical coherence, the principles sidestep

important questions such as what AI is, what constitutes a risk or benefit (and for whom), and what kind of futures various and diverse people might find desirable.

The Asilomar AI process, similar to the original Asilomar model, included a call for a moratorium on research, although the timing was significantly different in the Asilomar AI case. In the original Asilomar model, the goal of the moratorium was to provide time for experts to convene to discuss potential responses to possible biohazards associated with rDNA (Berg and Singer, 1995). In the Asilomar AI process, calls for a moratorium emerged five years following the initial release of the Asilomar AI principles. In March 2023, one week after the public release by Open AI of the large language model ChatGPT-4, the FLI issued an open letter calling for a six-month pause on the training of AI systems more powerful than GPT-4 (FLI, 2023). The rationale for the pause, according to the FLI letter, was to provide time for AI labs and experts to develop and implement safety protocols. The letter began with the following statement “AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research and acknowledged by top AI labs,” followed by a reference to the “widely-endorsed” Asilomar AI principles, and particularly the principle that “Advanced AI could represent a profound change in the history of life on Earth and should be planned for and managed with commensurate care and resources” (FLI, 2023). In the letter, the FLI lamented that current levels of planning and management had yet to adequately address the risks of AI. As such, a temporary pause was necessary, not for the development of AI in general, but to prevent a “dangerous race to ever-larger unpredictable black-box models with emergent capabilities” (FLI, 2023). The FLI claimed that AI development should be focused on ensuring that systems are developed in accordance with principles such as accuracy, safety, interpretability, transparency, robustness, alignment, trustworthiness, and loyalty. In the letter’s conclusion, the FLI stated that adherence to these principles can ensure that “humanity can enjoy a flourishing future with AI...(h)aving succeeded in creating powerful AI systems, we can now enjoy an AI summer: which we reap the rewards, engineer these systems for the clear benefit of all, and give society a chance to adapt” (FLI, 2023).

The FLI pause letter was well-received and to date has amassed over 30 000 signatures. However, the letter’s core assumption, that AI technologies will be developed and implemented with industry labs and developers at the helm in shaping ethical and regulatory pathways, was met with harsh criticism by AI ethicists and scientists. Immediately following the letter’s public release, Timnit Gebru, Emily Bender, Angelina McMillan-Major, and Margaret Mitchell issued a statement denouncing the letter’s misleading and fear-mongering rhetoric (Gebru et al., 2023). Terms used in the FLI pause letter, such as “powerful digital minds,” “human-competitive intelligence” and “artificial general intelligence,” alongside leading questions, such as “Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us,” oversell the capacities of automated computer systems, and, in so doing, deceive people by misattributing agency to synthetic media. This has the effect of shifting conversations about accountability and responsibility away from the companies that develop and profit from the commercialization of automated computer systems. Additionally, the pause letter positioned social actors as passive, where society is given a “chance to adapt” to AI rather than having opportunities to actively shape AI’s trajectory. In turn, calls for a pause are significantly at odds with the profit-seeking motives of the tech industry, making the prospects of implementation of an actual moratorium slim to none (see Richards et al., 2023).

4. Asilomar AI’s publicity

With the historical development of the Asilomar model as a backdrop, what are the implications for AI’s publicity of the FLI’s use of the Asilomar model? The Asilomar AI principles and process shape AI’s publicity in three interconnected ways: as an agenda-setting manoeuvre to promote longtermist beliefs; as an approach to policy debates that restricts public engagement; and as a mechanism to enhance industry interests and control of AI governance. We address each of these points in turn.

Promotion of longtermist beliefs: The FLI's longtermist perspective, which prioritizes the enhancement and maximization of humanity's potential in a distant future, is grounded in a concern with existential risk. No longer a fringe philosophy, longtermism is an influential policy agenda for AI governance that informs lawmakers and public concern about AI's future apocalyptic potential (McKelvey and Roberge, 2023). Longtermism provides a moral compass for well-endowed institutions such as the FLI, as well as Open Philanthropy, Open AI, and the now defunct Future of Humanities Institute (a sister institute to the FLI). As the coffers of longtermist organizations grow, so do their influence on policy and governance. The FLI's ability to secure millions of dollars in funding is due, in part, to longtermism adherents' uncritical belief in the transformative power of technology, and unwavering support of the political economic systems that generate the very problems its followers claim to address (Crary 2023, Naughton 2022). A key aspect of longtermist thinking is support for status quo institutions and practices: "longtermism calls on us to safeguard humanity's future in a manner that both diverts attention from current misery and leaves harmful socioeconomic structures critically unexamined" (Crary, 2023: 50).

Consistent with longtermist beliefs, the Asilomar AI principles promote both utopian and dystopian narratives about AI in ways that shift attention away from existing harms caused by automated systems and from the companies that create, endorse, and profit from these systems. Consider, for instance, differences between the FLI framing of its Asilomar AI principles and the original Asilomar framings of rDNA. Whereas the original Asilomar conferences focused on issues such as laboratory safety that were understood by scientific experts, the Asilomar AI conference focused on existential risks and doomsday speculations about abstract and largely unknown future dangers of artificial intelligence. Indeed, throughout the FLI's Asilomar AI principles, the phrase 'artificial intelligence' is invoked as if its meaning were self-evident yet its risks unknown. Whereas rDNA was a novel technology in the 1970s with yet-unrealized benefits and risks, autonomous computerized systems and algorithms are already deployed and existing harms have been well-documented by academic researchers, ethicists, and journalists, including labor issues (Williams et al., 2022), deepening inequality (Eubanks 2017), surveillance and privacy (Zuboff 2019), entrenched racism and sexism (Noble 2018, Buolamwini 2023), environmental degradation (Crawford 2021, Hogan 2015), and the consolidation of corporate power (Lewis-Kraus 2016), among numerous other issues, with solutions, policies, and frameworks available to address these problems.

Many high-profile AI researchers and tech developers follow the ideas set in place by Asilomar AI principles, and use public platforms to generate both fear and hype that AI will soon develop superhuman capabilities and pose existential risks to humans in the distant future. In an interview in *MIT Technology Review*, Max Tegmark, MIT professor and founder and president of FLI, stated that the goal of the FLI pause letter was, in part, to enable public conversations about the existential risks that AI poses to future generations (Heikkilä 2023). Two months after the FLI released its pause letter, the Center for AI Safety – along with Sam Altman of Open AI, Yoshua Bengio, and Geoffrey Hinton - released a short statement about the risks of rogue AI: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (Center for AI Safety, 2023). In the same month, Altman and colleagues wrote a blog calling for the "governance of superintelligence" (Altman et al., 2023). In line with longtermist rhetoric informing the Asilomar AI principles, debate focused on how to govern the eventual emergence of harmful AI to ensure its safety in the future, rather than on existing and well-known harms.

In addition to longtermist concerns about a distant future serving as a distraction from existing threats and very real risks, proponents of longtermist beliefs draw on and perpetuate discriminatory beliefs and attitudes, including racism, sexism, classism, and xenophobia, with roots in long-discredited eugenics ideologies (Gebru and Torres, 2024). Longtermism and its associated beliefs contribute the perpetuation

and exacerbation of social inequalities alongside the undermining of existing efforts to address ongoing injustices against marginalized communities.³

Limited opportunities for public engagement: Similar to the Asilomar model, Asilomar AI provided limited opportunities for participation beyond the circle of invited attendees at the Beneficial AI conference. Conference proceedings and principles were made publicly available, but diverse publics were not given the opportunity to shape the agenda or the content of discussion. Although it is well-known that current applications of AI disproportionately harm groups already disadvantaged by race, gender and socio-economic background, perspectives that account for these disadvantages were not represented in the Beneficial AI conference nor reflected in the Asilomar AI principles. Conspicuously absent from Beneficial AI conference, and from signatories on the Asilomar AI principles and FLI pause letter, were leading critical scholars and thought leaders on AI ethics such as Emily Bender, Meredith Broussard, Joy Buolamwini, Timnit Gebru, Meredith Whittaker, and Safiya Noble (to list only a few) and human-rights focused organizations such as the Distributed AI Research Institute (DAIR) (Prabhakaran et al, 2022). The insularity and lack of diversity in Asilomar AI is particularly problematic given that the FLI claims to speak about benefits of AI for all humanity yet does so from the limited perspective of an economically privileged, predominantly white constituency.

Entrenchment of industry interests and control: Whereas the original Asilomar was led by publicly-funded research scientists concerned about potential social implications of laboratory research, the Asilomar AI was coordinated by an interconnected group of wealthy people and well-funded technology institutions, all with vested financial interests in AI development. As Berg cautioned, the Asilomar model of scientific self-governance does not work if industry interests are at play:

...the best way to respond to concerns created by emerging knowledge or early-stage technologies is for scientists from publicly-funded institutions to find common cause with the wider public about the best way to regulate — as early as possible. Once scientists from corporations begin to dominate the research enterprise, it will simply be too late (Berg, 2008: 291).

Concern about the role of tech corporations in reshaping social and political orders at national and global scales has been a recurring theme of critical AI scholarship for some time (Zuboff, 2019; Dyer-Witford et al., 2019). In a review of AI ethics principles, Attard-Frost and colleagues concluded that important principles – such as fairness, accountability, and transparency – are undermined in practice by existing business norms centered on competitiveness and secrecy (Attard-Frost et al., 2023). The firing by Google of AI ethics researchers Timnit Gebru and Margaret Mitchell serves as a case in point of the inadequacy of well-meaning principles in transforming actual industry practices (Schiffer, 2021).

5. Beyond Asilomar

Since its inception five decades ago, the Asilomar model has taken on a symbolic role in science and technology policy. Often hailed as a success story of scientific solidarity in addressing social responsibility, Asilomar's legacy is seen in the privileged positioning of technical experts, particularly those with vested interests in the deployment of the technologies they develop, to render political judgements about how technology should be regulated and governed. The Asilomar model shapes technology regulation by narrowing the focus of public discussion to an examination of safety, impacts, and consequences, while sidelining broader questions about the purposes and imaginaries that animate technological research and innovation. Asilomar-style approaches, even revised ones that acknowledge the importance of diverse expertise, position publics and regulatory agencies as passive recipients of, and by-standers to, expert-driven self-governance.

³ Although beyond the scope of this paper to discuss in detail, the links between longtermism and eugenics ideologies warrant critical scrutiny, as do intersections between the ideologies driving discourses of safe A.I. and safe biotechnologies.

The FLI's Asilomar AI approach echoes and repeats many aspects of the original Asilomar model such as expert-driven principles, calls for a moratorium on research, and limited opportunities for public engagement. Unlike the original Asilomar's narrow focus on risks scientists knew best, the Asilomar AI model focused on existential risks for yet-to-be born humans, while downplaying known harms of current AI deployment. The FLI's Asilomar AI principles served to garner legitimacy for the longtermist agenda by linking AI products with a higher moral purpose, and by evoking Asilomar-style approaches to technology governance.

We see this paper as a contribution to growing literature on principle-based governance of AI. In contextualizing the Asilomar AI principles in a longer history, we draw connections with earlier efforts to govern biotechnology in the 1970s. Our concern rests primarily with the FLI's use of the Asilomar model to advance longtermist principles. We agree with critics who claim that the Asilomar model itself is a "poor and political dangerous model for governing emerging technologies" because of its limited scope and circumscribed opportunities for diverse participation (Parthasarathy, 2015: 305).

Here, we emphasize that it matters for AI's publicity how principles to govern AI are developed, by whom, and in line with which guiding ideas (Ferri and Gloerich 2023). Moving beyond the Asilomar model could entail proving opportunities to diversify engagement in the shaping of AI governance, yet engagement risks becoming a procedural end-in-itself if it serves as a surrogate and substitute for more radical visions and enactments of political judgement and politicized struggle over the future of AI (Barney, 2008). Enhancing democratic politics in AI's publicity should also entail directing attention to the purposes of AI research, development, and implementation; redistributing the profits of AI development from private interests to public commons; strengthening regulatory mechanisms to halt harmful applications of AI; and enacting broader institutional and structural changes to challenge existing power hierarchies and systems of oppression. We view the Asilomar AI principles as a harmful diversion that advances the misguided longtermist ideologies of the FLI and enhances the interests of those who stand to profit from the continued development of AI in the short term.

References

- ASOC (2010) The Asilomar Conference Recommendations on Principles for Research into Climate Engineering Techniques. Climate Institute. Washington, D.C. Available at: <https://climateviewer.com/downloads/Asilomar-Conference-Report-2010.pdf> (accessed December, 2023)
- Altman S, G B and Sutskever I (2023) Governance of Superintelligence. In: Open AI. Available at: <https://openai.com/blog/governance-of-superintelligence> (accessed December 1, 2023).
- Attard-Frost B, De los Ríos A and Walters DR (2023) The ethics of AI business practices: a review of 47 AI ethics guidelines. *AI and Ethics* 3(2): 389-406. <https://doi.org/10.1007/s43681-022-00156-6>
- Baltimore D, Berg P, Botchan M, et al. (2015) A prudent path forward for genomic engineering and germline gene modification. *Science* 348(6230): 36-38. <https://doi.org/10.12998/wjcc.v11.i3.528>
- Barinaga M (2000) Asilomar Revisited: Lessons for Today? *Science (American Association for the Advancement of Science)* 287(5458): 1584-1585. <https://doi.org/10.1126/science.287.5458.1584>
- Barney D (2008) Politics and emerging media: The revenge of publicity. *Global Media Journal -- Canadian Edition* 1(1): 89-106.
- Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada, pp.610 - 623. Association for Computing Machinery.
- Berg P (2004) *Asilomar and recombinant DNA*. Available at: <https://www.nobelprize.org/prizes/chemistry/1980/berg/article/>. (accessed December 1, 2023)
- Berg P (2006) Brilliant science, dark politics, uncertain law. *Jurimetrics* 46(4): 379-389. <http://www.jstor.org/stable/29762947>
- Berg P (2008) Asilomar 1975: DNA modification secured. *Nature* 455(7211): 290-291. <https://doi.org/10.1038/455290a>
- Berg P, Baltimore D, Boyer HW, et al. (1974) Biohazards of Recombinant DNA. *Science* 185: 3034.
- Berg P, Baltimore D, Brenner S, et al. (1975) Summary statement of the Asilomar Conference on recombinant DNA molecules. *Proc. Nat. Acad. Sci* 72: 1981-1984.
- Buolamwini J. (2023) *Unmasking AI: My mission to protect what is human in a world of machines*. New York: Random House.

- Bostrom N. (2014) *Superintelligence: Paths, dangers, strategies*. Oxford UK: Oxford University Press.
- Capron A (2015) The lessons of Asilomar for today's science. *New York Times*, May 28.
- Capron A and Shapiro R (2001) Remember Asilomar? Re-examining science's ethical and social responsibilities. *Perspectives in Biology and Medicine* 44(2): 162 - 169. <https://doi.org/10.1353/pbm.2001.0022>
- Center for AI Safety (2023) AI Extinction Statement Press Release. Available at: www.safe.ai/press-release (accessed December 1, 2023).
- Crary A (2023) The toxic ideology of longtermism. *Radical Philosophy* 214: 49 - 57.
- Crawford K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.
- Dyer-Witheford N, Mikkola Kjosien A and Steinhoff J (2019) *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. New York: Pluto Press.
- Editorial (2015) After Asilomar. *Nature* 526(7573): 293-294. <https://doi.org/10.1038/526293b>
- Eubanks V. (2017) *Automating inequality: how high-tech tools profile, police, and punish the poor*. New York: St. Martin's press.
- Ferber D (2004) Time for a Synthetic Biology Asilomar? *Science* 303(5655): 159-159. <https://doi.org/10.1126/science.303.5655.159>
- Ferri G, Gloerich I. (2023) Risk and harm: Unpacking ideologies in the AI discourse. *Proceedings of the 5th International conference on Conversational User Interfaces* 23: 1 - 6
- Fjeld J, Achten N, Hilligoss H, et al. (2020) Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center. January 15.
- Future of Life Institute (FLI) *Our mission*. Available at: <https://futureoflife.org/our-mission/> (accessed December 1, 2023).
- Future of Life Institute (FLI) (2017) *AI Principles*. Available at: <https://futureoflife.org/open-letter/ai-principles/> (accessed December 1, 2023).
- Future of Life Institute (FLI) (2023) *Pause giant AI experiments: An open letter*. Available at: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> (accessed December 1, 2023).
- Future of Life Institute (FLI) (n.d.) *Beneficial AI 2017*. Available at: <https://futureoflife.org/event/bai-2017/> (accessed December 1, 2023).
- Galaz V (2015) A manifesto for algorithms in the environment. *Guardian*, October 5. Available at: <https://www.theguardian.com/science/political-science/2015/oct/05/a-manifesto-for-algorithms-in-the-environment> (accessed December 1, 2023)
- Gebru T, Bender E, McMillan-Major A, et al. (2023) Statement from the listed authors of Stochastic Parrots on the "AI Pause" letter. DAIR. Available at: <https://www.dair-institute.org/blog/letter-statement-March2023/> (accessed December 1, 2023).
- Gebru T, Torres E. (2024). The TESCREL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday* 29 (4).
- Gottweis H (2005) Transnationalizing recombinant-DNA regulation: Between Asilomar, EMBO, the OECD, and the European Community. *Science as Culture* 14: 325 - 338. <https://doi.org/10.1080/09505430500369020>
- Grace K (2015) The Asilomar conference: A case study in risk mitigation. *Technical report 2015-9. Machine Intelligence Research Institute*. Berkeley, CA. Available at: <https://intelligence.org/files/TheAsilomarConference.pdf>
- Gregorowius D, Biller-Andorno N and Deplazes-Zemp A (2017) The role of scientific self-regulation for the control of genome editing in the human germline: The lessons from the Asilomar and the Napa meetings show how self-regulation and public deliberation can lead to regulation of new biotechnologies. *EMBO Rep* 18(3): 355-358. <https://doi.org/10.15252/embr.201643054>
- Heikkila M (2023). What's changed since the 'pause AI' letter six months ago? *MIT Technology Review* September 26. Available at: <https://www.technologyreview.com/2023/09/26/1080299/six-months-on-from-the-pause-letter/>. (accessed August 25, 2024).
- Hogan M (2015) Data flows and water woes: The Utah Data Center. *Big Data & Society* (July–December): 1–12 <https://doi.org/10.1177/2053951715592429>
- Hurlbut JB (2015a) Limits of Responsibility: Genome Editing, Asilomar, and the Politics of Deliberation. *Hastings Center Report* 45(5): 11-14.
- Hurlbut JB (2015b) Remembering the future: Science, law, and the legacy of Asilomar. In: Jasanoff S and Kim S-H (eds) *Dreamscapes of Moderity: Sociotechnical imaginaries and the fabrication of power*. Chicago, IL: University of Chicago Press, pp.126 - 151.
- Jackson DA, Symons RH and Berg P (1972) Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proc Natl Acad Sci U S A* 69(10): 2904-2909.
- Jasanoff S and Hurlbut JB (2018) A global observatory for gene editing. *Nature* 555: 435-437.
- Jasanoff S, Hurlbut JB and Saha K (2019) Democratic Governance of Human Germline Genome Editing. *Crispr j* 2(5): 266-271. <https://doi.org/10.1089/crispr.2019.0047>

- Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389-399. <https://doi.org/10.1038/s42256-019-0088-2>
- Krimsky S (1982) *Genetic Alchemy: The Social History of the Recombinant DNA Controversy*. Cambridge, Mas: MIT Press.
- Labaree A (2014) Our science-fiction apocalypse: Meet the scientists trying to predict the end of the world. *Salon*, May 10. Available at: <https://www.cser.ac.uk/news/our-science-fiction-apocalypse-meet-scientists-try/> (accessed August 25, 2024).
- Lewis-Kraus, G. (2016) The great A.I. awakening. *The New York Times Magazine*. Dec. 14. Available at: <https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html> (accessed August 25, 2024)
- MacAskill W (2022) *What we owe the future*. New York: Basic Books.
- Marchant G, Tournas L and Gutierrez CI (2020) Governing new technologies through soft law: Lessons for artificial intelligence. *Jurimetrics* 61(1): 1 - 18.
- McKelvey F and Roberge J (2023) Recursive Power: AI Governmentality and Technofutures. In: Lindgren S (ed) *Handbook of Critical Studies of Artificial Intelligence*. London: Elgar, pp.21 - 32.
- Milmo D (2023) Google, Microsoft, Open AI and start up form body to regulate AI development. *The Guardian*, July 16. Available at: <https://www.theguardian.com/technology/2023/jul/26/google-microsoft-openai-anthropic-ai-frontier-model-forum> (accessed August 25, 2024).
- Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 501–507
- Munn L (2023) The uselessness of AI ethics. *AI and Ethics* 3(3): 869-877. <https://doi.org/10.1007/s43681-022-00209-w>
- Naughton J. (2022) Longtermism: how good intentions and the rich created a dangerous creed. *The Guardian* Dec. 4. Available at: <https://www.theguardian.com/technology/commentisfree/2022/dec/04/longtermism-rich-effective-altruism-tech-dangerous>. (accessed August 25, 2024).
- Noble S. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York: NYU Press.
- Parthasarathy S (2015) Governance Lessons for CRISPR/Cas9 from the Missed Opportunities of Asilomar. *Ethics in Biology, Engineering and Medicine: An International Journal* 6: 305-312.
- Prabhakaran V, Mitchell M, Gebru T, Gabriel J. 2022. A human rights-based approach to responsible AI. *arXiv: 2210.02667* <https://doi.org/10.48550/arXiv.2210.02667>
- Richards B, Aguera Y, Arcas B, et al. (2023) The illusion of AI's Existential Risk. *Noema*, July 18. Available at: <https://www.noemamag.com/the-illusion-of-ais-existential-risk/> (accessed August 25, 2024).
- Rogers M (1975) The pandora's box congress. *Rolling Stone*, June 19. Available at: [https://web.mit.edu/endy/www/readings/RollingStone\(189\)37.pdf](https://web.mit.edu/endy/www/readings/RollingStone(189)37.pdf). (accessed December 1, 2023)
- Rufo F and Ficorilli A (2019) From Asilomar to Genome Editing: Research Ethics and Models of Decision. *NanoEthics* 13(3): 223-232.
- Schiffer Z (2021) Google fires second ethics researcher following internal investigation. *The Verge*, February 19. Available at: <https://www.theverge.com/2021/2/19/22292011/google-second-ethical-ai-researcher-fired>. (accessed December 1, 2023)
- Singer M and Soll D (1973) Guidelines for hybrid DNA molecules. *Science* 181: 1114.
- Stapf-Fine H, Bartosch U, Bauberger S, et al. (2018) Policy Paper on the Asilomar Principles on Artificial Intelligence. June. Berlin: Federation of German Scientists. Available at: https://vdw-ev.de/wp-content/uploads/2019/05/Policy-Paper-on-the-Asilomar-principles-on-Artificial-Intelligence_end.pdf (accessed December 1, 2023)
- Stirling B (2018) The Asilomar AI Principles. *Wired*, August 13. Available at: <https://www.wired.com/beyond-the-beyond/2018/06/asilomar-ai-principles/> (accessed December 1, 2023)
- Taylor C and Dewsbury B (2019) Barriers to inclusive deliberation and democratic governance of genetic technologies at the science-policy interface. *Journal of Science Communication* 18(3): Y02. <https://doi.org/10.22323/2.18030402>
- Torres É (2021) Against Longtermism. *Aeon*. Available at: <https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo>. (accessed December 1, 2023)
- Wagner B (2018) Ethics As An Escape From Regulation.: From “Ethics-washing” To Ethics-shopping? In: Bayamlioglu E, Baraliuc I, anssens J, et al. (eds) *Being Profiled*. Amsterdam: Amsterdam University Press, pp.84–89.
- Williams A, Miceli M, and Gebru T. (2022) The exploited labor behind artificial intelligence. *Noema*. <https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/>
- Zuboff S (2019) *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: The Hatchett Group.

X-gram and/as techspposure

Spelling out the climate consequences of generative AI

Crystal Chokshi

Mount Royal University, Canada

✉ cchokshi@mtroyal.ca

Abstract

How can generative AI's entanglement with climate crisis be made visible? This question catalyzed a research-creational intervention, called X-gram, that emblemizes this paper's theoretical contribution, a concept I call techspposure. Combining the words "technology" and "exposure," techspposure is a method by which a technology's affordances are made to expose its own material, infrastructural, environmental, or climate consequences. Put another way, techspposure occurs when tech tells on itself in a particularly poetically-just kind of way. Broadly, this article argues that research-creational interventions are urgently required ethical tools in our scholarly toolkits. They equip us to creatively destabilize Big Tech's charting of bleak climate futures and to engage audiences in imagining, as Natalie Loveless (2019) puts it, how the world could be organized differently.

Keywords: research-creation; techspposure; generative AI; climate crisis

1. Introduction

In 2021, journalist James Vincent, whose beat is artificial intelligence (AI), stated that "[l]anguage generation is the hottest thing in AI right now" (para. 1). The assertion was not a clever double-entendre. While Vincent went on to mention large language models' (LLM) problems with logic and (sexist and racist) language, his discussion did not address carbon output or climate impact. Yet at the time, a picture of just how "hot" generative AI really is was already emerging. Nine months earlier, Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell (writing under the wonderful pseudonym, Shmargaret Shmitchell) published a landmark paper, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" The paper connected climate crisis and generative AI, just as Emma Strubell et al. (2019) had done a couple of years earlier when they estimated that training one LLM produced the equivalent in carbon dioxide emissions as "125 round trip flights from NYC to Beijing" (Dobbe & Whittaker, 2019, para. 12). Not only did Bender et al.'s (2021) paper make similarly provocative claims, but it also took considerable "heat" for them: Google fired both Timnit Gebru and Margaret Mitchell, two of the paper's co-authors, in an attempt to discredit the work (Hao, 2020; Johnson, 2021; Schiffer 2021).

We can think of Gebru's and Mitchell's departure from Google as a watershed moment. Before their employment was terminated, Google, one of the biggest players in what is increasingly being referred to as the AI arms race (Chow & Perrigo, 2023; Johnson, 2019; Meacham, 2023; Roose, 2023), had people within its organization speaking about climate concerns and advocating for LLMs to be designed in a climate-conscious way. Generative AI's real "hotness"—its link to climate crisis—seemed to be under genuine scrutiny by Big Tech. But, at the moment of these employees' departure, this landscape changed.

The precise nature of this change came into relief by way of Google executive Jeff Dean's (2020) memo to the company about the "mistakes" Gebru and her co-authors had made: "[On the dangers of stochastic parrots"] ignored too much relevant research — for example, it talked about the environmental impact of large models, but disregarded subsequent research showing much greater efficiencies" (para. 7). Presumably, Dean's logic would hold that, if LLMs continue to grow while having less proportional impact, then their growth is sustainable. But there are two problems here, and they are intertwined: Dean is guilty of a red herring, and the red herring suggests we should accept what Max Liboiron (2021) calls a threshold theory of pollution. A threshold theory of pollution holds that the earth can withstand a certain amount of pollution before serious consequences occur in ecosystems and communities; in other words, a certain amount of pollution is acceptable (Liboiron, 2021). However, as many scholars have argued, there is no elsewhere, no "away" where this ostensibly acceptable amount of pollution resides without enacting climate harm (Davies, 2019; Liboiron, 2021; Tsing, 2015).

Nonetheless, pollution and climate harm are not the focus of current public discourse surrounding generative AI. In my lived experience as a Calgary-based post-secondary instructor and critical-media-oriented scholar who devoted the last four years to writing a dissertation about generative AI, I see and hear conversation attending chiefly to practice and pedagogy, not climate. While there is an occasional news story attempting to discuss AI's climate consequences, the narratives circulating at a far more rapid pace emphasize *how* we should be using and regulating generative AI in both industry and education, not *whether* or *if*.

Still, whether and if remain important questions. Critical media scholars and practitioners, who are on the front lines of knowledge-production about the entanglement between climate and technology, have a vital role to play in raising them. We have a role to play in calling out red herrings, calling into question ways of thinking that uphold the threshold theory of pollution, and disrupting the Big-Tech-led discourse that holds AI is getting more efficient and, therefore, less harmful. This is particularly pressing as the technology industry is expected by 2040 to contribute 14 per cent of global greenhouse gas emissions (Belkhir & Elmeligi, 2018), up from two to three per cent in 2021 (United Nations Environment Programme, 2021b).

Research-creation is one tool in our scholarly toolkits that equips us to creatively destabilize Big Tech's charting of a (bleak) future. In this article, I describe a research-creational project that aims to make visible generative AI's relationship with the climate, a relationship Big Tech and, now, the public discourse have downplayed. In doing so, I retain Natalie Loveless' (2019) conviction that research-creation has the power and potential to incite people "to care, and to care differently" (p. 107).

I proceed by making the following moves. I begin by saying more about research-creation, underscoring its emphasis on broadening the legibility of, and the horizon of response to, crisis. Then, I briefly trace several strategies deployed by Big Tech to render illegible AI's entanglement with climate breakdown. In response, I present a research-creational project called X-gram, a word-prediction program that brings climate crisis to the fore by predicting climate-crisis-related words only. I hold that X-gram is an illustration of a broader concept I call techspposure, a method by which a technology's affordances are made to expose its own material, infrastructural, environmental, or climate consequences. Ultimately, X-gram and techspposure draw attention to and subvert the values built into generative AI and, as Robin Mansell (2012) would say, our communication system writ large.

2. Research-creation: Choice method in times of crisis

First, I offer some clarity around the term “research-creation.” While other countries have adopted different names for research that has a significant artistic component, Canada has largely opted for “research-creation” (Chapman & Sawchuk, 2012; Loveless, 2019).¹ Canada’s Social Sciences and Humanities Research Council (2021) defines the term in the following way:

An approach to research that combines creative and academic research practices, and supports the development of knowledge and innovation through artistic expression, scholarly investigation, and experimentation. The creation process is situated within the research activity and produces critically informed work in a variety of media (art forms). (para. 24)

Notably, SSHRC’s definition emphasizes the fact that research-creation bridges art and research, but the exact specifications of that bridge are ill-defined.

This ill-defined relationship has not always landed easily in academia. Henk Borgdorff (2012) does not mince words as to why: “artistic research is not regarded here as ‘real’ research... or is seen as a lesser form of it” (p. 66). In this fraught context, Owen Chapman and Kim Sawchuk (2012) articulated four subcategories of research-creation, attempting to pinpoint the intersections at which “real” research would emerge from the research-creational process. The names of these subcategories—research-for-creation, research-from-creation, creative presentations of research, and creation-as-research—emphasize when research is actualized in the creative process. As such, I read the essay, published just over 10 years ago, as a legitimizing project, an attempt to “articulat[e] and elaborat[e] the field by developing distinctions” (Chapman & Sawchuk, 2012, p.13), insisting that research-creation has a valid place at our academic tables.

Several years later, our global context had shifted considerably and so, too, the conversation about research-creation. Natalie Loveless’ (2019) arguments in favor of research-creation not only took its legitimacy as a given, but they also pleaded a very strong case for its urgency:

There is no longer any question that we are living in compromised times... Global ecological and economic collapse are discussed with alarming regularity... The arts have an important and often overlooked part to play in this context. They offer modes of sensuous, aesthetic attunement, and work as a conduit to focus attention, elicit public discourse and shape cultural imaginaries. “How might the world be organized differently” is a question that matters urgently, and it is a question that art—particularly art attuned to human and more-than-human social justice—asks in generative and complex ways. (p. 16)

In this short passage, Loveless characterizes the contemporary moment as one of crisis and collapse. She argues that what is necessary in response are methods that can appropriately respond to it. An appropriate response is to explore alternatives—to ask the question, “How can the world be organized differently?”—and this is the unique capacity of arts-based projects. They have the capacity to captivate, or what Loveless calls “attune.” By Loveless’ account, it is characteristic of the art object to inspire conversation and, perhaps, new ideas in the public imagination.

If attunement is one part of this equation, then the other part is a deliberateness about whose attention is being attuned, and in what ways: “[r]esearch-creation asks, what output forms—modes of publication—might most interestingly and generatively render [our] research public?” (Loveless, 2019, p. 31). That is, research-creation is preoccupied not only with outputs but also with outcomes. The difference is that the former counts what the researcher has done; the latter counts what has shifted in the minds, attitudes, or behaviours of the viewer (Kelleher & Males, 2020).

The research-creational work I have encountered over the last several years, and which has shaped my thinking about research-creation, has been demonstrative of both preoccupations, attunement and audience, in relation to questions surrounding technologies and the natural, or more-than-human, world. I am thinking of Matt Parker’s (2014, 2017) sonospheric investigations into data centres, which capture

¹ It is worth noting that, even in Canada, “research-creation” goes by many other names, including “practice-based research, practice-led research, research-based practice, research-led practice, creative-praxis, arts-driven inquiry, arts-based research, and, increasingly, artistic research” (Loveless, 2019, p. 4).

their dramatic soundscapes; Zane Griffin Talley Cooper's (2018) immersive virtual reality film about cryptocurrency, *Alchemical Infrastructures: Making Blockchain in Iceland*, which raises questions about cryptocurrency's relations with the land; Felix Loftus' (2021) multimodal camera, called *Conjuring Landscapes*, which positions nature as photographer; the Low-Carbon Research Methods' lab's (2022) 'zine conference, which displaced human travel and centred creativity in knowledge mobilization; Brian Sutherland's (2022) creation of a prototype solar-powered work station; and Andrew Bateman's (2022) video project, *Landscapes of the Arctic from a Southern City*, which explores the entanglement of distant cities by way of migratory birds. Each of these works has turned to art to weave academically-grounded knowledge into projects that make crises and, in some cases, potential responses to them, legible and thinkable in new ways, ways that enfold the listener/reader/viewer into the very question that Loveless (2019) articulates: "How might the world be organized differently?" (p. 16).

In this way, research-creation is not just a legitimate academic practice. It is also an eminently ethical one. In climate breakdown, we no longer have the luxury of shying away from the question directly above or asking it in a way that is ultimately indifferent to responses. Broadening legibility and inciting people "to care, and to care differently" is a moral imperative, especially when we consider what we are up against.

3. What we are up against: Strategies of invisibility

In the context of this article, what we are "up against" are strategies deployed by Big Tech to render invisible AI's entanglement with climate breakdown. In this section, I briefly outline three.

3.1 Ethical AI movement's silence on climate

Sustainability is vastly underrepresented in the ethical AI movement (Jobin, Ienca, & Vayena, 2019). In a content analysis of "84 documents containing ethical principles or guidelines for AI," Anna Jobin et al. (2019) found that sustainability (including mentions of environment, energy, nature, and resources) was mentioned in only 14 of them, "suggest[ing] that these topics might be currently flying under the radar of the mainstream ethical discourse on AI" (p. 15). The authors argue that the "underrepresentation of sustainability-related principles is particularly problematic in light of the fact that the deployment of AI requires massive computational resources which, in turn, require high energy consumption" (p. 15). Arif Ali Khan et al.'s (2021) systematic literature review corroborated earlier findings, and it argued that the four most common ethical principles surrounding AI are transparency, privacy, accountability, and fairness.

A quick scan of the major AI players' AI principles (Table 1) shows no deviation from the conclusions drawn in Jobin et al.'s (2019) and Khan et al.'s (2021) work. For its part, Google's (n.d.) AI principles do not make mention of the environment, climate crisis, or energy. This is despite the fact that their first principle in AI development is "Be socially beneficial" (Pichai, 2018, para. 6). Microsoft is no better. Notwithstanding its multi-billion-dollar investment in OpenAI (Microsoft, 2023), which has created one of the largest LLMs in the world, its "6 principles of responsible AI" (Microsoft, n.d., para. 1) are silent on climate. Open AI's (2018) charter similarly emphasizes business-as-usual.

The troubling implication of these studies and these principles is that the ethical AI movement is not likely to prioritize climate crisis anytime soon. Neither are AI consumers, not least of all because the connection between AI and climate is largely invisible: "Virtual, ethereal, weightless, clean, futuristic: these and other terms like them are what first come to mind for most people when they think of digital technologies" (Deibert, 2020, p. 209). Thus, generative AI, to hearken back to the quote that opened this paper, can somehow be both the hottest and the coolest thing in technology right now.

Table 1. Google's, Microsoft's, and OpenAI's principles for developing AI

Google	Microsoft	OpenAI
Be socially beneficial. Avoid creating or reinforcing unfair bias. Be built and tested for safety. Be accountable to people. Incorporate privacy design principles. Uphold high standards of scientific excellence. Be made available for uses that accord with these principles.	Fairness Reliability and safety Privacy and security Inclusiveness Transparency Accountability	Broadly distributed benefits Long-term safety Technical leadership Cooperative orientation

3.2 Nature-based metaphors

Metaphor is a powerful strategy in upholding this paradox, as the language surrounding AI and other contemporary digital technologies often position them as what Sally Wyatt (2021) calls “nature-based” (p. 410)—that is, as naturally-occurring objects or resources while downplaying, or erasing, their realities as material, energy-intensive sites. For example, “[d]ominant data metaphors,” write Tim Hwang and Karen Levy (2015), “consistently compare data to naturally occurring physical resources” (para. 12), such as “liquid (data streams), a solid (data mining), or a gas (the cloud)” (para. 8). The cloud, in particular, reconceptualizes massive energy- and water-consuming data centres as natural objects known for both their weightlessness and life-giving properties (Hogan, 2015; Croker, 2020). And metaphor such as data ecologies (Hogan, 2018), digital ecosystems (Krivy, 2023), and organic search (Jobin & Ziewitz, 2018) call to mind forests and fruit rather than carbon and climate. Metaphors embedded in “AI” follow in this thematic footprint. Not least of all, “intelligence” likens AI to a human brain (Boucher, 2019; Wallenborn, 2022), which is one of the few things in the world capable of producing electricity without burning fossil fuels.

Scholars in the field of environmental media have long challenged the perception of technology as immaterial and climate-neutral. Their work has described the entanglements between technology that, on one hand, appears ephemeral and, on the other, has significant material realities and consequences (Cubitt, 2017; Hogan, 2015; Maxwell & Miller, 2012; Nakamura, 2011; Pasek, 2019; Starosielski & Walker, 2016). These projects have been vital in exposing just how “hot” tech can truly get.

3.3 Creation of ignorance

Critical perspectives are not mainstream in part because of the values encoded into the digital spaces in which we seek information. Jutta Haider and Malte Rödl (2023) argue that Google, for instance, enacts “a new ignorance logic” in relation to climate change. Search results, and the algorithms underlying them, subscribe to “a data capitalist ideology (West, 2019) and embody a logic that assumes end users as consumers (Mager, 2012)” (Haider & Rödl, 2023, p. 9). Their point can be illustrated by way of a simple Google search for “AI and climate change” (Figure 1). Results generated for this search embed particular values, namely, in this case, that “[e]xpectations for the future of AI technologies are great... and are entwined with our sense of progress and development” (White & Lidskog, 2022, p. 494). What is rendered invisible in the search, argue Haider and Rödl, are keywords, representing assumptions, that Google seems to insert into the search without the searcher’s knowing. In this case, “good,” “productive,” or “useful” seem invisibly to accompany the terms I input. In this way, Google determines which horizons we can see in relation to AI and climate change and which horizons we remain ignorant of.

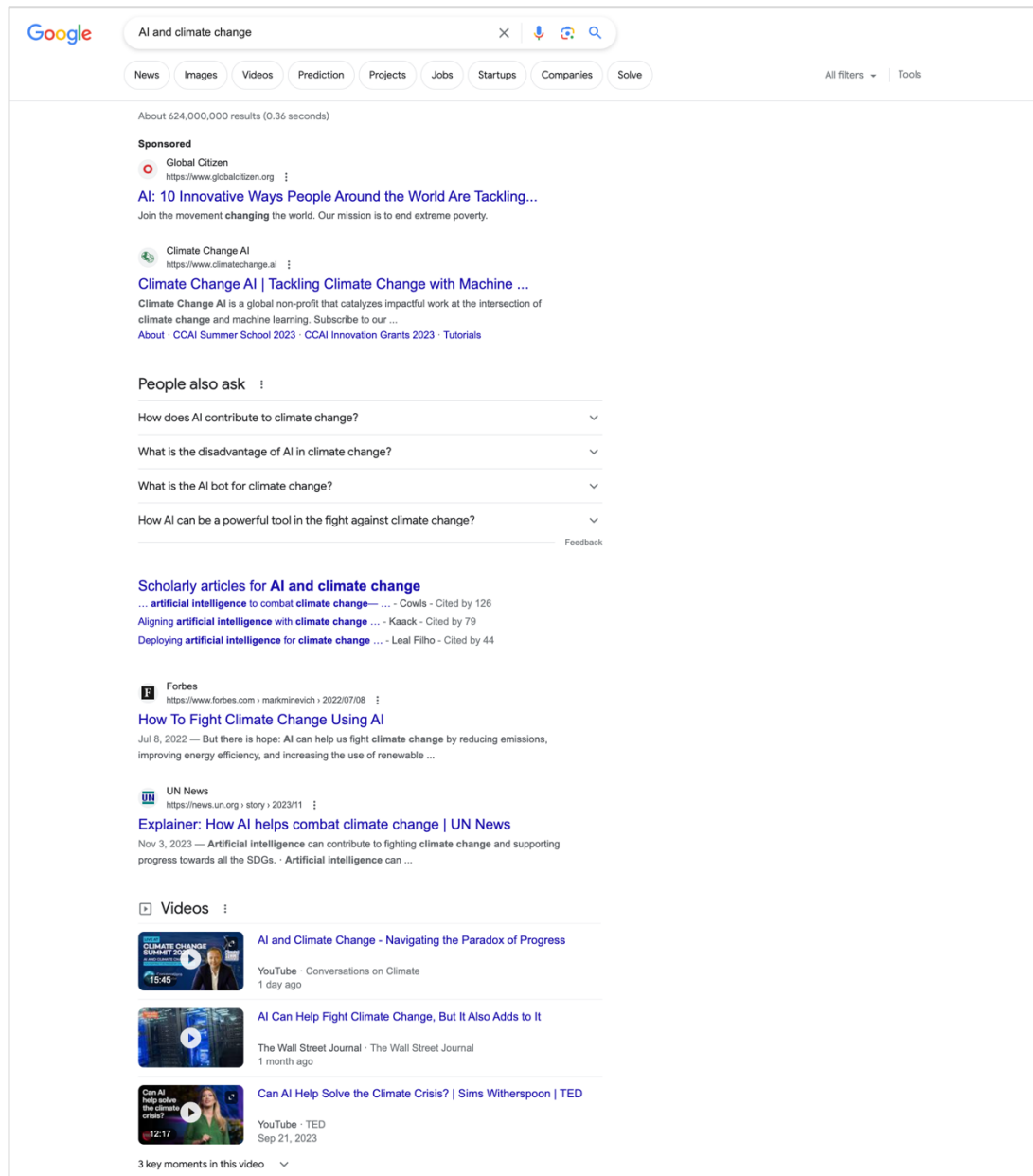


Figure 1. “AI and climate change”

A Google search for “AI and climate change” (conducted by the author in November 2023) yields results skewed to portraying AI as an ally in the fight against climate change

4. X-gram and/as techspposure

With AI anchored in Big-Tech-led efforts to render invisible its links with the climate crisis, I asked the following question: thinking with a research-creational mindset, how might AI’s entanglement with climate be made more visible? This question catalyzed the development of a research-creational intervention, called X-gram, and a wider concept, called techspposure, of which X-gram is illustrative. In this section, I describe X-gram, its theoretical underpinnings, and, finally, its relationship to techspposure.

It seemed to me that there could be no better way to demonstrate generative AI’s entanglement with the climate crisis than, quite literally, to spell it out. As such, I imagined a word-prediction program that would insist on predicting climate-crisis-related words only (Figure 2). This parameter defines X-gram.

Whereas popular generative AI applications, such as ChatGPT, have been trained on terabytes of data to make context-sensitive predictions, X-gram rejects context as the premise for predictive text. Ignoring narrative context, X-gram instead predicts, time and again, one of approximately 160 words related to climate crisis (Figure 3).²

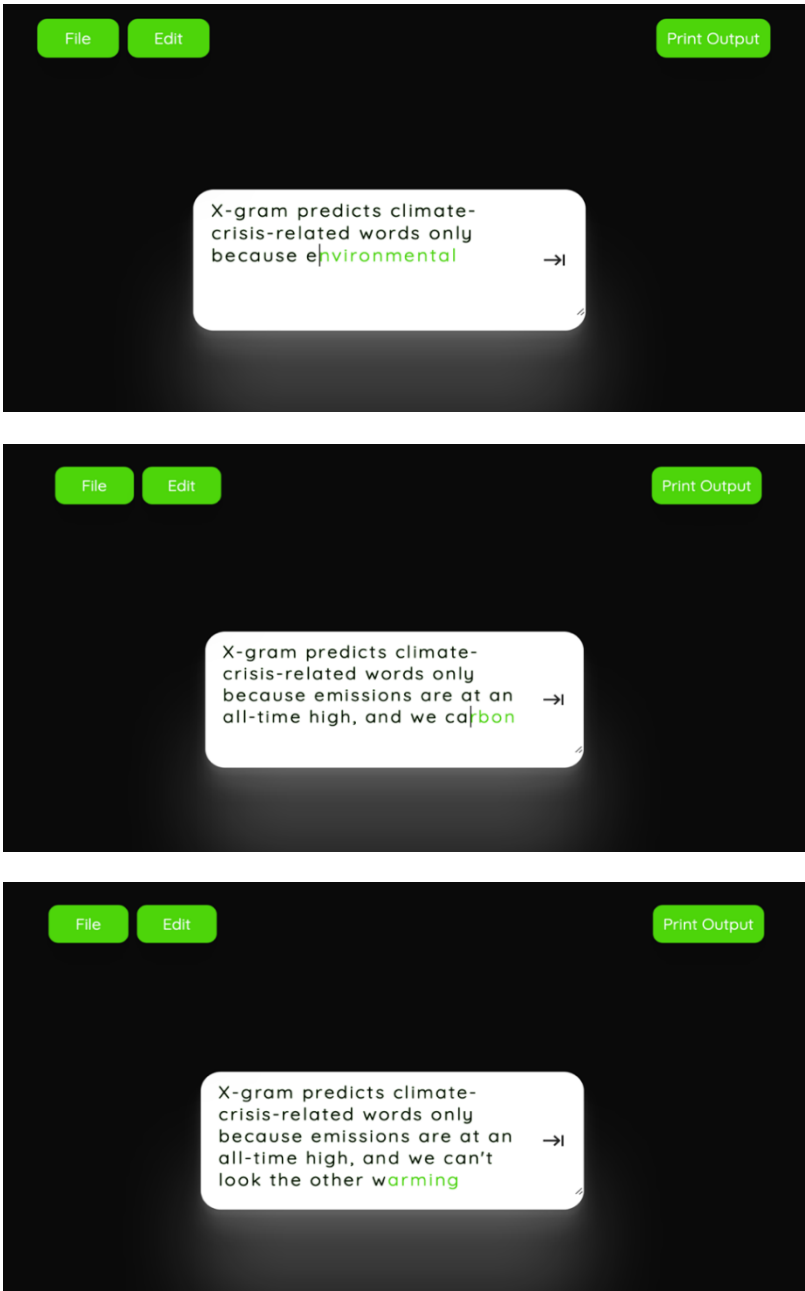


Figure 2. X-gram
X-gram describes its purpose

² X-gram’s development was possible thanks to Felix Loftus, a computational artist and technician based in London, UK. While I provided the research and specifications for X-gram, Felix coded X-gram into being.



Figure 3. X-gram ignores context
In this case, X-gram provides context-inappropriate text predictions for Johnny Nash’s (1972) song, “I can see clearly now”

Its limited predictive functionality means that X-gram is not generative AI. Nor does it aim to be, since instantiating AI would render X-gram complicit with the criticisms I lodge here. Rather, X-gram simulates generative AI. More specifically, it simulates Smart Compose, the generative AI Google has embedded in Workspace (a suite of applications represented by Gmail, Google Docs, Slides, Sheets, and Drawings, among others) since 2018. Unlike ChatGPT, which, following the input of a prompt, can

produce hundreds of words, Smart Compose predicts just one or two words at a time, allowing the writer to observe each act of prediction. When enabled, Smart Compose makes predictions for the next word(s) a writer might type, with predictions appearing in light grey text. To accept the prediction(s), the writer presses “tab.” Similarly, X-gram’s predictions appear in light green type, and writers accept predictions by pressing “tab,” just like Smart Compose.³

The words X-gram predicts originate from a corpus consisting of six texts: the Intergovernmental Panel on Climate Change’s (IPCC) homepage; the IPCC report titled AR6 Climate Change 2021: The Physical Science Basis (Summary for Policy Makers); the United Nations Environment Programme’s Emissions Gap Report 2021 (United Nations Environment Programme, 2021a); “Data flows and water woes: The Utah Data Center” (Hogan, 2015); “On the dangers of stochastic parrots: Can language models be too big?” (Bender et al., 2021); and “Pollution is colonialism” (CLEAR, 2017), a blog post based on Liboiron’s (2021) book by the same name. My motivation for choosing these particular texts was, to borrow words from Matthew Kirschenbaum (2016), imperfect, idiosyncratic, serendipitous, and duly diligent. Liboiron’s (2021) book deepened my understanding of the relationship between colonialism and the climate crisis. Bender et al.’s (2021) paper represents, as aforementioned, a watershed moment, changing the terms of conversations about climate and tech. Mél Hogan (2015) introduced me to the environmental turn in media studies, now firmly anchored as “environmental media,” and the impact of her work was visible to me, not least of all, when I taught her article in an upper-level undergraduate course: students reported that it exposed them to material realities they had not been aware of, and it changed their thinking more than any other text encountered in the course. As for the IPCC homepage, the IPCC report, and the UN report, these are texts that have become incredibly urgent. Including them was a means of recognizing their urgency and thinking toward new means to circulate them in public consciousness.

Using the web-based text-analysis tool called Voyant, I queried each text for its top 30 terms (see Figure 4 for an example). Then, I made the following eliminations. First, closely related terms in a single text were removed. For example, “change” and “changes” both appeared in AR6; therefore, I removed “changes.” Second, I removed terms relating to the medium or media wherein the works were published (e.g., “http” and “doi.org”). In the end, X-gram was programmed to predict a total of 163 words.

As aforementioned, X-gram’s prediction capability is deliberately limited. For example, when “c” is input by the writer, X-gram will predict one of the words beginning with “c” in its word bank; when “a” is input, X-gram will predict one of the words in its bank beginning with “a,” and so on. As such, the program is not dependent on LLMs or AI. Nor does it aspire to be. Even though I have introduced X-gram as a word-prediction program, predictive text is not the game at which X-gram is playing.

Rather, X-gram draws on other theoretical-methodological traditions to bring about something else altogether. The first of its influences is what Jacob Gaboury (2018) calls critical unmaking toward queer computation. This method “foreground[s] queer techniques of refusal, misuse, and disruption that must nonetheless work with and through contemporary digital technologies” (p. 484). Importantly, Gaboury is not suggesting that we invent radically new technologies in order to carry out this practice. Rather, he says critical unmaking happens with and through contemporary—i.e., existing—digital technologies (admittedly, however, his use of “nonetheless” in the quote above suggests resignation, not necessarily hopefulness, in this constraint). Specifically, critical unmaking happens when technologies are refused, misused, or disrupted in particular ways that demonstrate their neoliberal logics and values:

To compute queerly, then, is to acknowledge, embrace, and enact a practice of radical technological failure. It is to engage in critical unmaking: to make central those externalities—exploits, bugs, breakdown, abuse, and misuse—of our digital culture that, while pervasive, we nonetheless disavow... In acknowledging, accepting, and even producing failure, queer computation seeks to make clear the values and assumptions that drive our culture of technological development and to offer alternate modes of living with and through technology. (p. 485).

³ X-gram is available online for viewing and interaction at [X-gram.glitch.me](https://x-gram.glitch.me).

In sum, critical unmaking toward queer computation challenges its practitioners to foreground a politics of failure in order to lay bare, question, and ultimately reframe the ways we think about technology and the ways we think technology is “productive.”

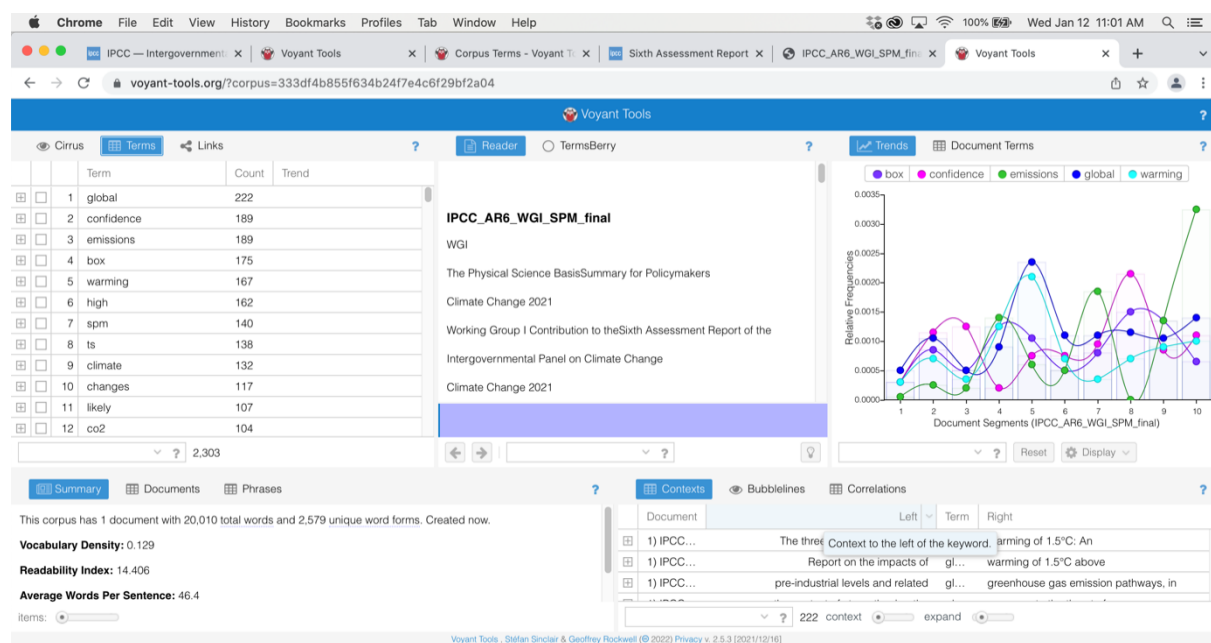


Figure 4. Voyant

A screen shot of the text analysis that Voyant performed on the IPCC report titled “AR6 Climate Change 2021”

X-gram takes this provocation seriously. In response, the program deploys techniques of refusal, misuse, and disruption that Gaboury calls for, and it deploys them in—or more accurately, weaponizes them against—existing contemporary technologies. X-gram’s interface mimics a rudimentary word processor, and its behaviours invite writers to think of it as having predictive-text AI. Most notably, just like Smart Compose, X-gram makes predictions in a lighter type that writers can accept by pressing “tab” (Figure 5). But just as quickly as X-gram orients writers to it as a certain kind of technology, the program alienates them by defying conventions around, and expectations of, what constitutes appropriate prediction logic. Where conventional word-prediction models are always scanning and memorizing text in order to make context-sensitive predictions, X-gram rejects context as the only, or the most legitimate, premise for word prediction (as evidenced in Figure 3). X-gram’s “File,” “Edit,” and “Print” buttons also yield unexpected results. “File” links writers to the six texts that were text-mined for X-gram’s word bank (Figure 6). Clicking on “Edit” reveals a broken menu full of short statements beginning with “Can’t”: “Can’t undo,” “Can’t redo,” “Can’t copy,” “Can’t paste,” “Can’t clear” (Figure 7). Lastly, clicking “Print” does not print the text a writer inputs. Rather, “Print” yields a PDF showing the full list of words in X-gram’s word bank.

In these ways, X-gram draws on the idea of the glitch. A glitch is “an aberration” (Gaboury, 2018, p. 485) or an “aesthetics of error” (Schneider, 2021, p. 261) in digital media. A glitch is an occurrence that is unexpected and conventionally unproductive, either in time or function. As Schneider (2021) puts it, “Glitches *arrest* us. They hold us up” (p. 262). That is, they prevent users from doing what users expect to do. In this way, Michael Betancourt (2016) argues, a glitch also plays an “interruptive role, suggest[ing] the materiality of media, offering the potential for a transition into a critique of digital capitalism” (p. 8), echoing Schneider’s notion that glitch art can “promote critical inquiry into and against the functional norms of capitalism” (p. 262). Indeed, in disrupting the conventional functionality of both word-

prediction AI and word processors, X-gram holds writers up, inviting them to pause on the consequences of AI-assisted writing. In doing so, X-gram embodies glitch logics and aesthetics.

Through critical unmaking and glitch, X-gram becomes what Luke Stark (2014) calls a data visceralization. Data visceralizations transform data into affective experience. They differ from visualizations insofar as they “are representations of information that don’t rely solely and primarily on sight or sound, but on multiple senses including touch, smell, and even taste, working together to stimulate our feelings as well as our thoughts” (para. 3). While X-gram involves sight, it is also very much a tactile experience. Writers must engage with touch via a keyboard in order to experience both the data at X-gram’s core (the six climate texts) and an affective response to them. It is also an experience likely to generate frustration and confusion. X-gram’s refusal to adhere to design imperatives characteristic of word-prediction AI, efficiency and speed, is likely to prompt writers to ask: what is this? And what is it trying to get me to do? If these questions surface, X-gram will have lived up to what Stark (2014) describes as the purpose of data visceralization work: “Why does making data more visceral matter? Simply put, we’re too often disconnected from the information we put out into cyberspace” (para. 14). X-gram’s purpose is to connect AI and climate change, and then to connect writers with this connection.

Drawing on critical unmaking, glitch, and data visceralization, X-gram makes this connection in a way that brings about something else altogether. I call it *techspposure*. Combining the words “technology” and “exposure,” techspposure is a method by which a technology’s affordances are made to expose its own consequences. Put another way, techspposure occurs when a technology tells on itself in a particularly poetically-just kind of way. X-gram does this by turning word-prediction tech into tech that insists on predicting *only* words that suggest its link to climate crisis. In doing so, X-gram not only connects writers to data from six important climate texts, but it also announces its own performative function. X-gram’s potential is in its capacity to shift writers’ perceptions of what we are engaged in—of what we are *doing*—when we use generative AI.

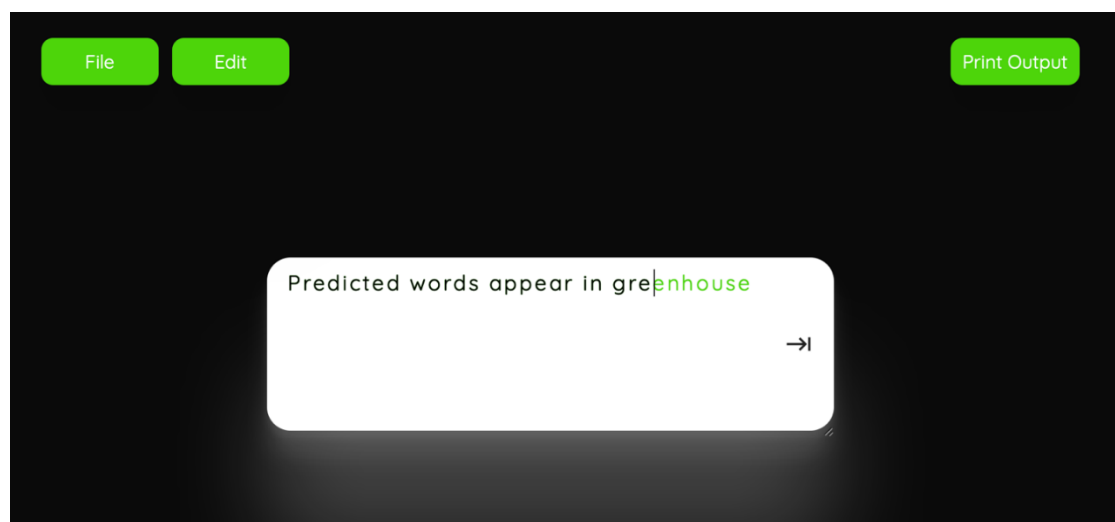


Figure 5. X-gram’s borrows user-interface elements from Google’s Smart Compose
X-gram’s predictions appear in light green type, and writers accept predictions by pressing “tab”

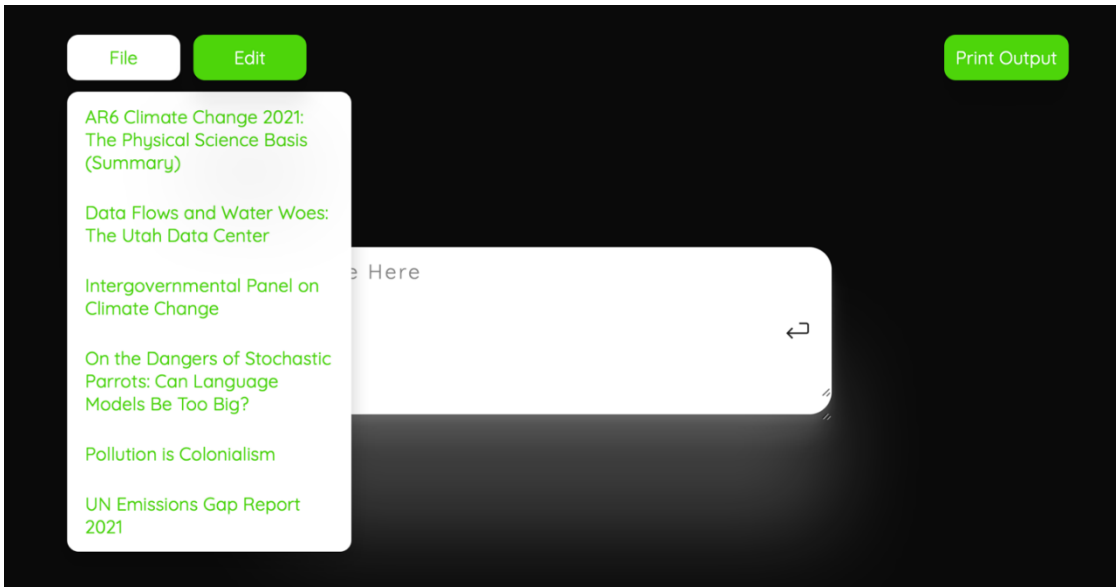


Figure 6. X-gram’s “File” menu
X-gram’s “File” menu provides links to the six texts that were mined for X-gram’s word bank

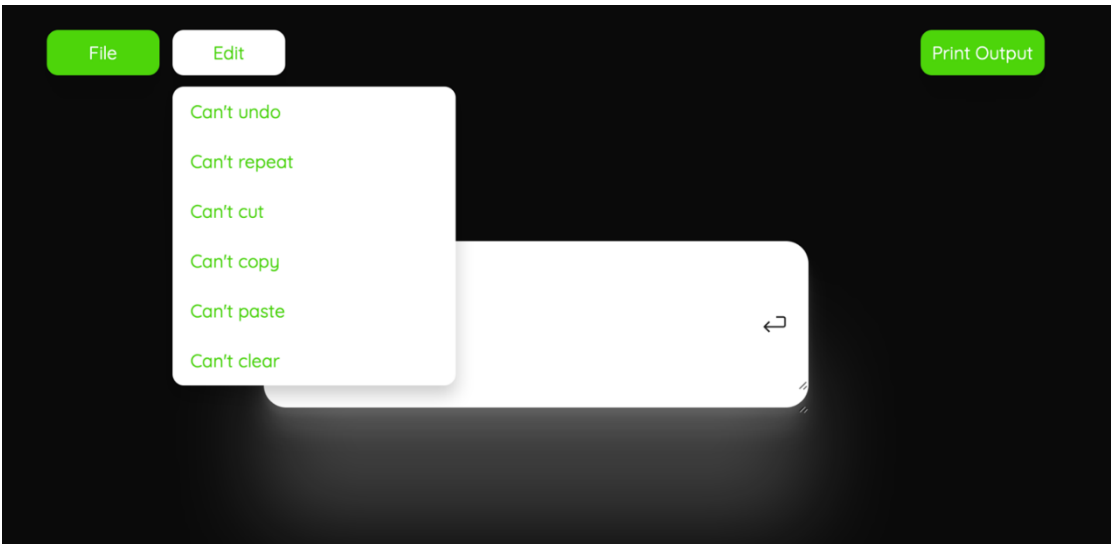


Figure 7. X-gram’s “Edit” menu
X-gram’s “Edit” menu reveals a list of “Can’t” statements

5. Related work

Other research efforts have been conceptualized to link more explicitly LLMs, AI, and climate, but not in the vein of techspouse. ClimateBERT, for example, is a LLM whose purpose is to train applications to better predict climate-related words (Webersinke, Kraus, Bingler, & Leippold, 2022). Because ClimateBERT was trained on climate-related texts, deploying the LLM means that climate-related words will be better predicted than, say, when deploying a LLM that has not received specialized training. ClimaText, for its part, is “a dataset for sentence-based climate change topic detection” (Varini et al., 2021, p. 1). It aims to scan large volumes of text for content relating to climate and, more than that, to carry out several actions on that content, such as sentiment analysis (Varini et al., 2021). A project by

Jonathan T.K. Chan (2021) aims to make climate texts more legible to the general public. Chan's "NLP Project - Text Summarization of IPCC Report" succeeded in reducing the reading difficulty level of the IPCC's Summary for Policymakers (2018). Finally, *The Climate Does Not Exist* shows the impacts of climate crisis, "one address at a time" (Tousignant, 2021, para. 3). The project imagines various locations around the world having been impacted by dimensions of climate change, and it generates images showing locations under these new conditions.

While making links between generative AI and the climate, the projects outlined above nonetheless reify the opacity of AI's entanglement with the climate. By contrast, X-gram, one instantiation of techspouse, prototypes a scenario in which one of generative AI's affordances, predictive text, is the very thing that reveals its material consequences.

6. Future work

Today, X-gram is largely a prototype. Available for viewing and interaction at [X-gram.glitch.me](https://x-gram.glitch.me), users will see the application's scope and functionality are limited—glitchy, even. It is early days in X-gram's journey.

But even so, X-gram is working toward research-creation's two hallmarks, audience and attunement, seeking generatively to engage audiences in the questions, how is the world organized? And, how might it be organized differently?

Which audiences does X-gram seek to engage? Planned future work will begin with students who will be invited to expand X-gram's scope and also to think through examples of techspouse. Subsequently, conference workshops will pursue the same.

What might other examples of techspouse look like? Imagine opening the photos app on a smartphone and seeing photographs that document mining practices involved in camera production. Find My iPhone could be reconfigured not to pinpoint the current location of my phone but rather to show the supply chains through which my phone passed to land in my hands. Or, as Ron Deibert (2020) muses about natural resources mined for tech production, "I have often thought how wonderful it would be if devices had a list of ingredients in the way that packaged food items do" (p. 214); I imagine clicking "About This Mac" on my computer and seeing the list Deibert has conceived. Projects like these would redeploy, with poetic justice, technological affordances so that their consequences are visible and clear.

The point here, too, is research-creation. While collaborating on the production of low-fidelity prototypes (which, with sufficient funding, will become interactive public exhibits) workshop participants will experience creation-as-research, one of the four subcategories of research-creation Chapman and Sawchuk (2012) elucidate. "'Creation-as-research' involves the elaboration of projects where creation is required in order for research to emerge" (p. 19): in other words, creation, and the emergence of research questions during creation, establish a new research agenda. My own work in process with X-gram has generated many questions, and I certainly look forward to the questions that participants and I will co-create.

Still, most exciting to me are, simply, the workshops themselves—and the possibility that readers of this article may organize their own. After all, neither X-gram nor techspouse are about taking on Big Tech. They are, rather, a simple invitation to put aside the neoliberal values surrounding generative AI and digital technologies, more broadly; to consider the often invisible relationships and consequences engendered by digital technologies; and to displace *how* in favor of *whether* and *if*.

Acknowledgements

This research was funded by the Social Sciences and Humanities Research Council as well as the Killam Trusts. I am grateful to Dr. Mél Hogan, Dr. Robin Mansell, Dr. Annie Rudd, Dr. Pamela Banting, the late

Dr. Halcyon Lawrence, and the anonymous peer reviewers who provided invaluable comments about this work.

References

- Bateman, A. (2022, April 20). Landscapes of the Arctic from a southern city. *Heliotrope*.
<https://www.heliotropejournal.net/helio/landscapes-of-the-arctic>
- Belkhir, L., & Elmeligi, A. (2018). Assessing ICT global emissions footprint: Trends to 2040 & recommendations. *Journal of Cleaner Production*, 177, 448-463. <https://doi.org/10.1016/j.jclepro.2017.12.239>
- Bender, E.M., Gebru, T. McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery*, 610-623. <https://doi.org/10.1145/3442188.3445922>
- Betancourt, M. (2016). *Glitch art in theory and practice*. Routledge.
- Borgdorff, H. (2012). *The conflict of the faculties: Perspectives on artistic research and academia*. Leiden University Press.
- Boucher, P. (2021, July 2). *What if we chose new metaphors for artificial intelligence?* European Parliamentary Research Service. <https://epthinktank.eu/2021/07/02/what-if-we-chose-new-metaphors-for-artificial-intelligence-science-and-technology-podcast/>
- Chan, J.T.K. (2021). *NLP Project - Text Summarization of IPCC Report*. https://github.com/jon-tk-chan/ipcc_nlp_analysis
- Chapman, O., & Sawchuk, K. (2012). Research-creation: Intervention, analysis and 'family resemblances.' *Canadian Journal of Communication*, 37(1), 5-26. <https://doi.org/10.22230/cjc.2012v37n1a2489>
- Chow, A.R., & Perrigo, B. (2023, February 17). The AI arms race is changing everything. *Time*. <https://time.com/6255952/ai-impact-chatgpt-microsoft-google/>
- CLEAR. (2017, September 1). *Pollution is colonialism*. <https://civiclaboratory.nl/2017/09/01/pollution-is-colonialism/>
- Cooper, Z.G.T. (2020). *Alchemical infrastructures: Making blockchain in Iceland* [Video].
<https://www.alchemicalinfrastructures.com/exhibit/spaces>
- Crocker, T.D. (2020). *Formation of the cloud: History, metaphor, and materiality* [Unpublished doctoral dissertation]. Virginia Tech.
- Cubitt, S. (2017). *Finite media*. Duke University Press.
- Davies, T. (2019). Slow violence and toxic geographies: "Out of sight" to whom? *Environment and Planning C: Politics and Space*, 40(2), 409-427. <https://doi.org/10.1177/2399654419841063>
- Dean, J. (2020). About Google's approach to research publication.
<https://docs.google.com/document/d/1f2kYWDXwhzYnq8ebVtuk9CqQqz7ScqxhSIxeYGrWjK0/edit>
- Deibert, R.J. (2020). *Reset: Reclaiming the internet for civil society*. Anansi.
- Dobbe, R., & Whittaker, M. (2019, October 17). *AI and climate change: How they're connected, and what we can do about it*. AI Now. <https://ainowinstitute.org/publication/ai-and-climate-change-how-theyre-connected-and-what-we-can-do-about-it>
- Gaboury, J. (2018). Critical unmaking: Toward a queer computation. In J. Sayers (Ed.), *The Routledge companion to media studies and digital humanities* (pp. 483-491). Routledge.
- Google AI. (n.d.). *Artificial intelligence at Google: Our principles*. <https://ai.google/principles/>
- Haider, J., & Rödl, M. (2023). Google Search and the creation of ignorance: The case of climate crisis. *Big Data & Society*, 10(1), 1-12. <https://doi.org/10.1177/20539517231158997>
- Hao, K. (2020, December 4). *We read the paper that forced Timnit Gebru out of Google. Here's what it says*. MIT Technology Review. <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>
- Hogan, M. (2018). Big data ecologies. *Ephemera*, 18(3), 631-657.
<https://ephemerajournal.org/sites/default/files/pdfs/contribution/18-3hogan.pdf>
- Hogan, M. (2015). Data flows and water woes: The Utah Data Center. *Big Data & Society*, 2(2), 1-12.
<https://doi.org/10.1177/2053951715592429>
- Hwang, T., & Levy, K. (2015, January 15). 'The cloud' and other dangerous metaphors. *The Atlantic*.
<https://www.theatlantic.com/technology/archive/2015/01/the-cloud-and-other-dangerous-metaphors/384518/>
- Intergovernmental Panel on Climate Change. (2023). *The intergovernmental panel on climate change*. <https://www.ipcc.ch/>
- Intergovernmental Panel on Climate Change. (2021). AR6 climate change 2021: The physical science basis (summary for policy makers). https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_SPM.pdf
- Jobin, A., Ienca, M., & Vayena, E. (2019). *The global landscape of AI ethics guidelines*. arXiv.
<https://doi.org/10.48550/arXiv.1906.11668>
- Jobin, A., & Ziewitz, M. (2018, March 6). Organic search: How metaphors help cultivate the web. *Digital Society Blog*.
<https://www.hiig.de/en/organic-search-metaphors-help-cultivate-web/>
- Johnson, K. (2021, April 10). *Black women, AI, and overcoming historical patterns of abuse*. VentureBeat. <https://www.media.mit.edu/articles/black-women-ai-and-overcoming-historical-patterns-of-abuse/>
- Johnson, J. (2019, October 29). *Is the US losing the artificial intelligence arms race?* The Conversation.
<https://theconversation.com/is-the-us-losing-the-artificial-intelligence-arms-race-124969>

- Kelleher, T., & Males, A.M. (2020). *Public relations* [Canadian edition]. Oxford University Press.
- Khan, A.A., Badshah, S., Liang, P., Khan, B., Waseem, M., Niazi, M., & Akbar, M.A. (2021). *Ethics of AI: A systematic literature review of principles and challenges*. arXiv. <https://doi.org/10.48550/arXiv.2109.07906>
- Kirschenbaum, M. (2016). *Track changes: A literary history of word processing*. Belknap Press of Harvard University Press.
- Krivy, M. (2023). Digital ecosystem: The journey of a metaphor. *Digital Geography and Society*, 5, 1-9. <https://doi.org/10.1016/j.diggeo.2023.100057>
- Liboiron, M. (2021). *Pollution is colonialism*. Duke University Press.
- Loftus, F. (2021). *Conjuring landscapes* [Multimedia project]. <https://floftus.glitch.me/conjuring.html>
- Loveless, N. (2019). *How to make art at the end of the world: A manifesto for research-creation*. Duke University Press.
- Low-Carbon Research Methods (2022). *DIY methods 2022*. <https://www.diymethods.net/>
- Mansell, R. (2012). *Imagining the internet: Communication, innovation, and governance*. Oxford University Press.
- Maxwell, R., & Miller, T. (2012). *Greening the media*. Oxford University Press.
- Meacham, S. (2023, September 8). A race to extinction: How great power competition is making artificial intelligence existentially dangerous. *Harvard International Review*. <https://hir.harvard.edu/a-race-to-extinction-how-great-power-competition-is-making-artificial-intelligence-existentially-dangerous/>
- Microsoft (2023, January 23). *Microsoft and OpenAI extend partnership*. <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/>
- Microsoft. (n.d.). *Principles and approach*. <https://www.microsoft.com/en-us/ai/principles-and-approach>
- Nakamura, L. (2011). Economies of digital production in East Asia: iPhone girls and the transnational circuits of cool. *Media Fields Journal*, 2. <http://mediafieldsjournal.org/economies-of-digital/>
- OpenAI. (2018, April 9). *OpenAI charter*. <https://openai.com/charter>
- Parker, M. (2017). *The people's cloud* [Video series]. <https://thepeoplescloud.org/>
- Parker, M. (2014). *The cloud is more than air and water* [Audiovisual installation]. <https://www.earthkeptwarm.com/the-cloud-is-more-than-air-and-water/>
- Pasek, A. (2019). Managing carbon and data flows: Fungible forms of mediation in the cloud. *Culture Machine*, 18. <https://culturemachine.net/vol-18-the-nature-of-data-centers/managing-carbon/>
- Pichai, S. (2018, June 7). *AI at Google: Our principles*. <https://blog.google/technology/ai/ai-principles/>
- Roose, K. (2023, February 3). How ChatGPT kicked off an AI arms race. *The New York Times*. <https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html>
- Schiffer, Z. (2021, March 5). *Timnit Gebru was fired from Google—then the harassers arrived*. The Verge. <https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean>
- Schneider, R. (2021). Glitch. In Thylstrup, N., Agostinho, D., Ring, A., D'Ignazio, C., & Veel, K. (Eds.), *Uncertain archives: Critical keywords for big data* (pp. 259-270). MIT Press.
- Social Sciences and Humanities Research Council. (2021, May 4). *Definition of terms*. <https://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/definitions-eng.aspx#a22>
- Stark, L. (2014, May 19). Come on feel the data (and smell it). *The Atlantic*. <https://www.theatlantic.com/technology/archive/2014/05/data-visceralization/370899/>
- Starosielski, N., & Walker, J., (2016). *Sustainable media: Critical approaches to media and environment*. Routledge.
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. <https://doi.org/10.18653/v1/P19-1355>
- Sutherland, B. (2022). Strategies for degrowth computing. *LIMITS '22: Workshop on Computing within Limits*. <https://computingwithinlimits.org/2022/papers/limits22-final-Sutherland.pdf>
- Tousignant, B. (2021, October 13). *The climate does not exist: Picturing impacts of the climate crisis with AI, one address at a time*. <https://mila.quebec/en/article/this-climate-does-not-exist-picturing-impacts-of-the-climate-crisis-with-ai-one-address-at-a-time/>
- Tsing, A.L. (2015). *The mushroom at the end of the world: On the possibility of life in capitalist ruins*. Princeton University Press.
- United Nations Environment Programme. (2021a). *Emissions gap report 2021: The heat is on*. United Nations Environment Programme and United Nations Environment Programme Copenhagen Climate Centre. <https://www.unep.org/resources/emissions-gap-report-2021>
- United Nations Environment Programme. (2021b, March 19). *With new pact, tech companies take on climate change*. <https://www.unep.org/news-and-stories/story/new-pact-tech-companies-take-climate-change>
- Varini, F.S., Boyd-Graber, J., Ciaramita, M., Leppold, M. (2021). ClimaText: A dataset for climate change topic detection. arXiv. <https://doi.org/10.48550/arXiv.2012.00483>
- Vincent, J. (2021, December 8). *DeepMind tests the limits of large AI language systems with 280-billion-parameter model*. The Verge. <https://www.theverge.com/2021/12/8/22822199/large-language-models-ai-deepmind-scaling-gopher>
- Wallenborn, J.T. (2022, May 17). AI as a flying blue brain? How metaphors influence our visions of AI. *Digital Society Blog*. <https://www.hiig.de/en/ai-metaphors/>

- Webersinke, N., Kraus, M., Bingler, J.A., & Leippold, M. (2022). ClimateBERT: A pretrained language model for climate-related text. arXiv. <https://arxiv.org/pdf/2110.12010.pdf>
- White, J.M., & Lidskog, R. (2022). Ignorance and the regulation of artificial intelligence. *Journal of Risk Research*, 25(4), 488-500. <https://doi.org/10.1080/13669877.2021.1957985>
- Wyatt, S. (2021). Metaphors in critical Internet and digital media studies. *New Media & Society*, 23(2), 406–416. <https://doi.org/10.1177/1461444820929324>