Feminist Philosophy Quarterly

Volume 8 | Issue 3/4 Introduction

2022

Feminism, Social Justice, and Artificial Intelligence

Carla Fehr https://orcid.org/0000-0002-4533-1589
University of Waterloo
carla.fehr@uwaterloo.ca

Recommended Citation
Fehr, Carla. 2022. “Feminism, Social Justice, and Artificial Intelligence.” *Feminist Philosophy Quarterly* 8 (3/4). Introduction.

Feminism, Social Justice, and Artificial Intelligence[1]

Carla Fehr

Artificial Intelligence (AI) profoundly affects issues of justice and well-being in individual, social, and global contexts. From social media to search engines, and in domains ranging from policing and judicial decision making, to the assessment of insurance, university, and job applications, to the creation of visual arts and beyond, algorithms are embedded in many of our lives. Some of these algorithms promise incredible contributions to human welfare. For example, some AI-powered systems can detect very early stages of medical conditions, and others are being developed and used to combat human trafficking. Also, consider social media's role in liberatory social movements such as Arab Spring, Black Lives Matter, and #MeToo.

However, many developments in AI have significant adverse impacts, including job loss, privacy violations, political polarization, and the spread of disinformation. Scholars such as Safiya Noble, Cathy O'Neil, and Ruha Benjamin have demonstrated the pervasive, real-world negative consequences of algorithmic bias and discrimination against racialized people, women, and members of other marginalized groups. Noble (2018) recounts the insult of using "black girls" as a keyword in a Google search for activities for her stepdaughter and nieces and being primarily directed to pornography sites. Noble documents the phenomenon of *algorithmic oppression*, the "masking and deepening of social inequality" resulting from discrimination against racialized people and women that is "embedded in computer code, and increasingly, in artificial intelligence technologies that we are reliant on, by choice or not" (Noble 2018, 1). She argues that "algorithmic oppression is not just a glitch in the system but, rather, is fundamental to the operating system of the web" (Noble 2018, 10). Benjamin (2019) refers to the near-ubiquitous technologies that amplify racial hierarchies as the New Jim Code. O'Neil (2016) coins the term "weapons of math destruction" (WMD) to describe algorithms that do harm (for example, encoding racism), impact many people, and function as a black box

[1] Many thanks to Laura Foster, Katy Fulfer, Jesse Hoey, Catherine Hundleby, Trystan Goetze, Leah Govia, Aimée Morrison, Kem-Laurin Lubin, Lynne Sargent, Jamie Sewell, and twenty-seven anonymous reviewers. Your labour and expertise were vital to the production of this special issue. I would also like to thank the Feminism, Social Justice and AI workshop participants and the authors of the papers in this volume. It is comforting and inspiring to be part of a community of scholars who generously supported each other's work and conduct a kind of philosophy that makes the world a better place.

blocking the inner workings of the algorithm from evaluation. O'Neil provides a host of examples of WMDs that, under a thin veneer of objectivity, reify and recreate cultural biases that systematically harm members of marginalized groups. One example describes recidivism models that overpredict future criminal acts of Black defendants and underpredict future criminal acts of white defendants.

Given these justice-focused benefits and harms of AI, AI continues to be an apt and urgent topic of feminist philosophical engagement.

This special issue is the culmination of a collaborative effort that began with the 2021 Feminism, Social Justice, and AI workshop. Workshop participants were solicited through a public call for proposals. Selected participants submitted full papers, which we discussed and developed. Participants were then invited to revise papers and submit them for consideration for this issue. The papers in this issue were drawn from that pool of submissions after a double-anonymous review.

Understanding and Addressing Algorithmic Bias and Discrimination

One group of papers in this volume focuses on understanding and addressing algorithmic bias and discrimination. Within this group are papers focusing on the *barriers or challenges to debiasing algorithms*. Oisín Deery and Katherine Bailey characterize a dilemma inherent in debiasing some algorithms. They point out that when assessing algorithms, there is a trade-off between normative correctness and descriptive correctness. They use the example of a Google search for images of CEOs. Only about 5 percent of Fortune 500 companies have women CEOs. So, should the output of Google search results reflect this number or a number representing a more equitable situation? Deery and Bailey write that in cases like this, two options present themselves.

> Either prioritize descriptive accuracy over normative correctness, which has the potential cost of perpetuating or amplifying bias, or instead prioritize normative correctness, with the potential cost of withholding ethically useful information, even if we thereby avoid perpetuating or amplifying bias. (19)

Deery and Bailey argue that we will likely need to make these choices on a case-by-case basis. In many situations, the preferred solution will be to control for the adverse effects of bias rather than debiasing the models themselves. Linus Ta-Lun Huang, Hsiang-Yun Chen, Ying-Tung Lin, Tsung-Ren Huang, and Tzu-Wei Hung focus on the challenges of debiasing explainable AI (XAI). XAI models have more transparent decision-making processes and are more likely to expose their biases than other AI systems. Huang, Chen, Lin, Huang, and Hung explain that technical XAI, the view that technical experts can handle debiasing XAI, is mistaken. Instead, they advance and

advocate for integrated XAI, which draws on diverse and marginalized perspectives in developing and assessing XAI.

The second subgroup of papers on algorithmic bias and discrimination focuses on *bias and fairness in terms of structural injustice.* Using health care as a case study, Ting-An Lin and Po-Hsuan Cameron Chen frame AI bias as a structural injustice. Lin and Chen argue that fairness cannot be achieved by computational means alone because there is a need to address social structure and power imbalances in AI development and use. Deploying Iris Marion Young's social connection model, Lin and Chen argue for distributing the responsibility for AI-mediated injustices among those who participate in the injustice, and they provide a set of practical recommendations for the pursuit of AI fairness. Alysha Kassam and Patricia Marino confront the problem of proxy discrimination, which arises when an algorithm does not consider sensitive characteristics (such as race) but does consider putatively neutral characteristics (such as zip code) that turn out to be correlated with a sensitive characteristic. Proxy discrimination is a source of anti-Black racism. Kassam and Marino argue that fairness-as-parity, which aims at creating "equal rates of accurate and inaccurate predictions" between groups and is a common response to proxy discrimination, fails to address structural racism (2). Starting from a structural view of racism Kassam and Marino argue that "algorithms should be evaluated with respect to their broader social impact and whether their use exacerbates or mitigates racial stratification" (2).

Harms Perpetuated by AI

Several papers in this volume focus on understanding and addressing concrete harms created by AI. Emma McClure and Benjamin Wald use the example of Google searches to demonstrate that machine learning algorithms can communicate hostility and exclusion and so inflict environmental microaggressions on members of marginalized populations. They argue that tech companies such as Google should be proactive by retraining their algorithms on less biased datasets—for example, on Black Lives Matter archives—and should restrain their algorithms from doing harm by hiring people with lived experience to curate liberatory autocomplete responses for common racialized queries. Michael Randall Barnes explores the role of AI in online radicalization leading to real-world violence and argues that "better AI" is not a solution to this problem. He is concerned that in favouring a technical fix, "Big Tech reveals an overall ideology in which technological 'progress' is valued over human flourishing" (3). Barnes argues that these AI-centric solutions are a form of propaganda because a focus

> on the future *potential* of algorithmic solutions dehumanize [content moderators], obfuscate the harms they face, and complicate the

> debate about the distribution of responsibility in actually addressing these challenges. (3)

Barnes argues that regular users should pressure Big Tech to address this problem and that content moderators should be supported and included in the development and implementation of strategies for reducing online radicalization.

Epistemic Oppression and Injustice, and Algorithmic Oppression

Some papers in this volume demonstrate how algorithmic oppression contributes to epistemic oppression and injustice. A common expectation that racialized people explain the racism they experience or defend their assessments of experiences as racist can result in uncompensated and onerous labour that Nora Berenstain (2016) calls epistemic exploitation. Tempest M. Henning evaluates one strategy for resisting this epistemic exploitation: to suggest that one's interlocutors "just Google it." Henning considers how theories of argumentation would reject this response as evading the burden of proof and failing to engage in a collaborative argumentation project. She argues that these rejections of the "just Google it" strategy are unacceptable because they fail to consider the heavy cost inflicted on racialized people by requiring a defence of their experiences of racism. However, Henning rejects the "just Google it" strategy because of racism baked into Google search results. Not only is it unlikely that antiracist information would be suggested by a privileged or racist person's search, but they would also likely receive information that would make the problem worse. Given that epistemic exploitation is harmful and turning to Google is not viable, Henning considers alternative strategies, ranging from asking for compensation for the pedagogical labour to declining to engage in conversations about racism, for avoiding epistemic exploitation. Henning shows that racism built into AI blocks its use as an antiracist educational and political tool.

Heather Stewart, Emily Cichocki, and Carolyn McLeod argue that social media algorithms support and develop algorithmic sorting and targeting. Algorithmic sorting refers to grouping social media users into separate, closed-off, and biased informational communities (9), which can decrease exposure to people who hold different perspectives. Algorithmic targeting occurs when information is presented to a user based on predictions about what they want to see (9). Algorithmic targeting can decrease people's engagement with perspectives and information that challenge their assumptions and beliefs. Stewart, Cichocki, and McLeod argue that algorithmic sorting and targeting lead to social distrust and undermine cooperation, which is associated with members of marginalized communities being denied full status as knowers.

Arianna Falbo and Travis LaCroix argue the importance of investigating cultural code-switching in emerging AI technologies. Cultural code-switching refers to a

person changing how they present and construct themselves in response to changes in their social environment. While cultural code-switching can signal group membership, it can also lead to cultural smothering, a form of self-censoring in which "one alters aspects of their cultural identity in response to an unwelcoming or hostile social atmosphere" (3). Cultural smothering can harm members of marginalized groups. Falbo and LaCroix point out that we exist in relationship with AI systems that encode conventions of dominant cultures. As a result, cultural smothering can be mediated by AI systems. Falbo and LaCroix warn that a failure to implement code-switching capacities in AI risks entrenching and widening social inequalities.

Friction and Discomfort with Feminine and Colonial AI

Papers by Alexis Elder and Shelley M. Park demonstrate that it can be OK to feel uncomfortable in the face of AI that trades on misogyny and colonialism. Alexis Elder uses the Confucian moral concept *li* (禮) to understand and address the gendered abuse of feminized AI, such as Siri and Alexa, used as home assistants. Elder uses *li* in a way that refers to ritual or etiquette and "as a tool for resisting inherited habits and maladaptive patterns" (18). She writes, "These devices fail, not because they introduce or cause sexism but because they make it harder to resist the sexism that is already the water we swim in . . . . We need more friction when it comes to our assumptions and 'instinctive' actions around gender" (18).

In her analysis of social robots who do care work, Shelley Park offers a reinterpretation of the uncanny valley, which refers to human discomfort with human replicas that look almost but not quite human. The uncanniness portrayed in cultural depictions of social robots has become both an engineering and a marketing problem for robot engineers. Park recasts uncanniness in a psychoanalytic frame, arguing that social robots doing care work echo gendered and colonial labour and, as such, are a moral rather than a technological challenge. Park argues that our discomfort with uncanniness is an apt response to gendered colonial violence. She writes that "social justice may depend—in part—on designing robots that heighten rather than reduce our sense of the uncanny, leaving us less 'at home' with our intimate relationship to AI" (23).

Looking Forward

What would it mean to create feminist AI? While it remains vital that feminist scholars respond to harms perpetuated by extant AI systems, Os Keyes and Kathleen A. Creel demonstrate that opportunities remain for feminist scholarship on the development of new algorithmic systems. Keyes and Creel take inspiration from feminist philosopher Alison Adam's 1998 book, *Artificial Knowing: Gender and the Thinking Machine*, and revive her arguments in the context of current technology. Echoing Adam, Keyes and Creel imagine a feminist AI and "thinking through the ways

in which AI research could be informed by feminist theory" (Adam 1998, 156; quoted in Keyes and Creel, 3). For example, both the texts by Adam and by Keyes and Creel attend to whose knowledge and interests are represented in AI systems. Adam raised concerns about the unacknowledged situatedness of the *data included in* early AI systems, Cyc and Soar. Keyes and Creel point out current problems of the partiality and situatedness of the *data being used to train* current AI. Even though current systems are based on different technology, Keyes and Creel point out that the problem of AI being trained on biased data sets remains urgent. Finally, Keyes and Creel explore strategies for increasing the "representationality and plurality of machine learning systems' underlying 'knowers'" (3).

Cross-Cutting Themes

There are additional themes that cut across these groups of papers. First, several papers in this volume take a structural approach to feminism and AI and, in doing so, draw on the work of Iris Marion Young. In addition to the papers by Lin and Chen and by Kassam and Marino, Young's structural approach to oppression, as well as her social connection model of political responsibility, surfaces in the papers by Barnes; Falbo and LaCroix; and Stewart, Cichocki and McLeod. Second, papers by Lin and Chen; Barnes; Henning; and McClure and Wald, in addition to developing theoretical, philosophical positions, also support feminist praxis by providing concrete suggestions for addressing the problems they document. A third theme involves the importance of moving beyond technological solutions to problems of algorithmic bias and injustice in AI. This theme arises in the work of Barnes; McClure and Wald; Lin and Chen; Falbo and LaCroix; and Huang, Chen, Lin, Huang, and Hung.

Even though this issue focuses on feminism, social justice, and AI, these papers also engage a wide range of additional philosophical fields, including ethical and social theory, philosophy of science, epistemology, philosophy of language, psychoanalytic theory, and Confucian philosophy, demonstrating that feminism, social justice, and AI is a topic of broad philosophical interest.

References

Adam, Alison. 1998. *Artificial Knowing: Gender and the Thinking Machine*. New York: Routledge.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.

Berenstain, Nora. 2016. "Epistemic Exploitation." *Ergo* 3 (22) 569–90. https://doi.org/10.3998/ergo.12405314.0003.022.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

CARLA FEHR is the Wolfe Chair in Scientific and Technological Literacy in the Department of Philosophy at the University of Waterloo, where she works in socially responsible philosophy of science, feminist philosophy of science, feminist philosophy of biology, and feminist epistemology. Fehr is associate director of the APA-CSW Site Visit Program. She is a coeditor of *Feminist Philosophy Quarterly* and the editor of this "Feminism, Social Justice, and Artificial Intelligence" special issue.