Centre for International Governance Innovation

Digital Policy Hub

Digital Policy Hub – Working Paper

Avoiding Catastrophe Through Intersectionality in Global AI Governance

Laine McCrory

Fall 2024 cohort

About the Hub

The Digital Policy Hub at CIGI is a collaborative space for emerging scholars and innovative thinkers from the social, natural and applied sciences. It provides opportunities for undergraduate and graduate students and post-doctoral and visiting fellows to share and develop research on the rapid evolution and governance of transformative technologies. The Hub is founded on transdisciplinary approaches that seek to increase understanding of the socio-economic and technological impacts of digitalization and improve the quality and relevance of related research. Core research areas include data, economy and society; artificial intelligence; outer space; digitalization, security and democracy; and the environment and natural resources.

The Digital Policy Hub working papers are the product of research related to the Hub's identified themes prepared by participants during their fellowship.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

Partners

Thank you to Mitacs for its partnership and support of Digital Policy Hub fellows through the Accelerate program. We would also like to acknowledge the many universities, governments and private sector partners for their involvement allowing CIGI to offer this holistic research environment.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Key Points

- Artificial intelligence (AI) safety is a growing field that highlights the existential risks of AI, while proposing alternative development processes centred around concepts of alignment with human values and ethical concerns.
- While it promotes critical perspectives, AI safety has been criticized for its limited conceptualization of future existential threats as universally impactful. Critics of the AI safety movement have highlighted that the movement does not acknowledge how current groups are already experiencing disproportionate existential risks.
- This working paper utilizes a feminist policy analysis framework centred around five thematic areas — intersectionality, context, neutrality, control and power — to analyze global initiatives for AI safety governance.
- The analysis reveals that AI safety policies often lack meaningful engagement with feminist principles, failing to acknowledge how future risks are tied to current harms. As AI systems regularly replicate broader social biases and power dynamics, it is important to address how an intersectional perspective views future impacts as distinctly connected to current harms being faced by marginalized groups.
- Future AI safety work can benefit from integrating feminist perspectives such as accountability and participation into the research and policy development processes.

Introduction: Defining AI Safety

As AI systems develop at a rapid pace, AI safety has emerged as a discipline concerned with addressing the existential threats of AI through the development of systems aligned with human values such as ethics, explainability and external control (Amodei et al. 2016; Bostrom 2014). This movement has become increasingly prominent surrounding the utopian promises and dystopian risks of artificial general intelligence (AGI), described as "a single system that can learn incrementally, reason abstractly, and act effectively over a wide range of domains" (Voss 2017). Similar to the potential for global nuclear annihilation, AI safety experts warn of the potential for AGI to pose irreversible consequences to all human life. In 2002, Nick Bostrom argued that existential risk represents a global terminal threat that "would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" (Bostrom 2002, 2). Dan Hendrycks, Mantas Mazeika and Thomas Woodside (2023) identify four categories of AI existential risk: malicious use, a global unregulated AI race, organizational risks associated with a lack of safety culture and the potential for AGI to produce rogue agents. These scholars represent a growing body of literature focused on how the development of superintelligence could result in permanent harms (Bostrom 2014; Ord 2020; Vold and Harris 2021). In addition to within academic research, numerous well-known figures in the field have voiced concerns for the future of AI, including Stephen Hawking (Cellan-Jones 2014) and Bill Gates (Eadicicco 2015).

Globally, AI safety has gained importance within both policy and research and development sectors, with the founding of AI safety institutes in the United States, United Kingdom, Japan, Canada, Singapore and the European Union (Variengien and Martinet 2024). These institutes supplement industry commitments to develop safe AI, with Cohere, OpenAI and AI21 agreeing on best practices to mitigate misuse (Cohere Team 2022), industry leaders

signing onto pledges that acknowledge the potential existential threats of AI[1] and calls for a pause on giant AI experiments (Future of Life Institute 2023). In 2023, UN Secretary-General António Guterres advocated for a new UN agency to govern the potential catastrophic risks of AI (Fung 2023). While the pathways for addressing the risks associated with AGI are diverse (Sotala and Yampolskiy 2014), they represent an interdisciplinary movement seeking to collaborate on existential challenges.

Safety on Whose Terms?

Yet, AI safety proponents often face criticism for a research agenda that is too forward looking, neglecting the current harms experienced by marginalized groups (Acemoglu 2021). Scholarship in critical data studies demonstrates how AI reinforces inequalities related to race, sexuality, class and gender (Buolamwini and Gebru 2018; Bender et al. 2021; Hamidi, Scheuerman and Branham 2018), creating environmental (Crawford 2021), racial (Benjamin 2019), gendered (D'Ignazio and Klein 2020) and neocolonial (Tacheva and Ramasubramanian 2023) impacts. Subsequently, for Timnit Gebru and Émile Torres (2024), the values underscoring the development of AGI — utopian aspirations and industry commitments to safety — reinforce oppressive conditions. They argue that the field promotes universalized notions of safety that fail to hold developers accountable for the practices of marginalization, extraction and exploitation that form the basis of these models.

AI research is characterized by a diversity problem, facing not only a lack of diverse perspectives within technical development (Stinson and Vlaad 2024), but also a lack of representation in the framing of policy narratives and ethical principles (Roche, Wall and Lewis 2023; Ulnicane and Aden 2023). When there is an ongoing lack of diverse perspectives, the narratives surrounding what constitutes an existential threat become insular. This lack of diversity means that any approach to AI safety will not adequately address the questions of participation, accountability and representation (Lazar and Nelson 2023).

Method: Constructing a Feminist AI Policy Framework

This research is grounded in a feminist account of the hegemonic impacts of AI. In reiterating Beverly McPhail's (2003) notion of policy as a phenomenon that shapes the world around us, this paper proposes a policy framework that anchors future AI safety efforts in interdisciplinarity and intersectionality.

Key to this work is a theoretical approach that reflects an intersectional depiction of existential AI threats. Coined by Kimberlé Crenshaw (1991), intersectionality is a framework that analyzes how social identities (gender, race, ethnicity, sexual identity, gender identity, class, religion, national origin, documentation status, migration status, carceral status and ability) compound and overlap, creating different forms of discrimination and/or privilege. By addressing how compounding identities become systematically marginalized, an intersectional analysis examines how systemic injustices are reified in daily life.

Other frameworks propose alternative critiques and constructions of AI and policy. For Catherine D'Ignazio and Lauren Klein (2020; 2024), the concept of "data feminism"

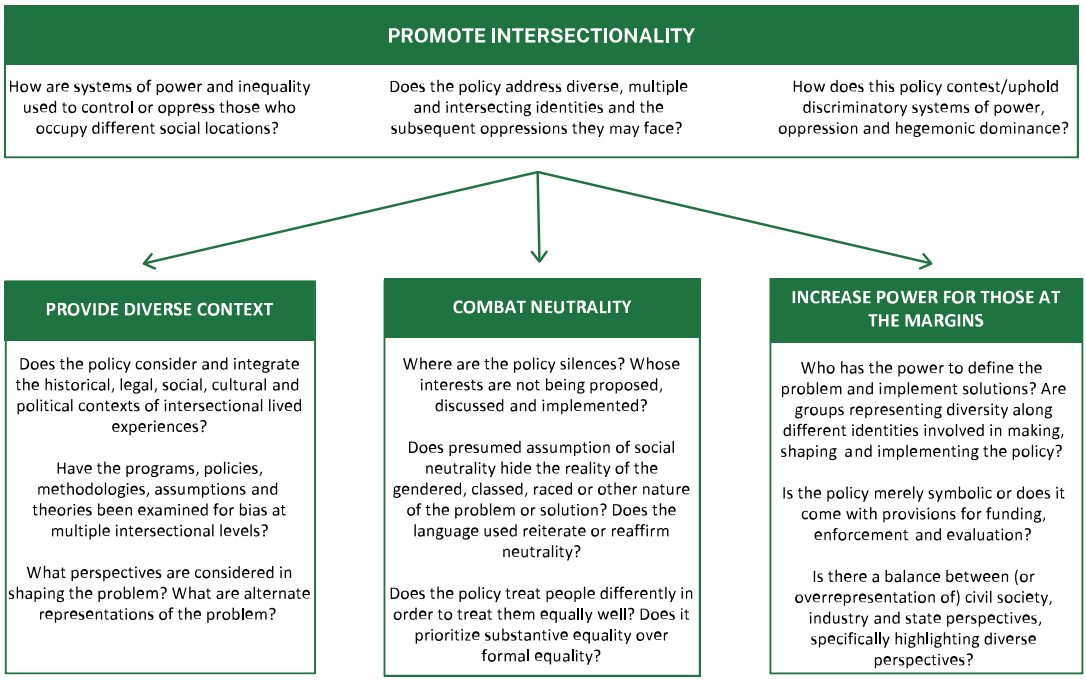
1 See www.safe.ai/work/statement-on-ai-risk#open-letter.

represents a framework for integrating feminist principles into the development of data and AI, including examining and challenging power, rethinking hierarchies and considering contexts. In 2003, McPhail devised the Feminist Policy Analysis Framework, a qualitative action-oriented model that utilizes 13 categories to examine how public policy reifies gendered systems of oppression. In 2020, Heather Kanenberg, Roberta Leal and Stephen Erich amended McPhail's framework with considerations of how policy decisions have intersectional impacts. They argue that feminist policy analysis reveals marginalizing and discriminatory practices, connecting them to broader socio-political concerns (Kanenberg, Leal and Erich 2020). Approaches to feminist policy analysis also exist within social policy (Hyde 2000; Hankivsky and Cormier 2011), education (Mansfield, Welton and Grogan 2014; Bensimon and Marshall 2003), public health (Hankivsky et al. 2014) and social work (Kanenberg 2013). These frameworks are united in their commitment to critically reviewing how power intersects with policy in uneven and marginalizing ways.

Drawing from these frameworks, this research proposes a feminist AI policy framework (see Figure 1), which encourages decision makers and stakeholders to evaluate potential AI safety projects in accordance with four goals. Within each of these goals is a series of questions that can be asked when examining AI safety initiatives.

- **Promote intersectionality:** Growing from approaches to feminist AI governance that reiterate the need to see AI as a product of structural inequalities, colonial legacies and disproportionately marginalizing harms (Crawford 2021; Ricaurte 2024; Toupin 2024), a focus on intersectionality acts as an umbrella goal by examining how power, context and neutrality impact different groups according to their social identity.
- **Provide diverse contexts:** Assessing the context of a policy involves bringing hidden narratives to light. Focusing on the contexts that influence these intersectional identities — histories, backgrounds, social structures — shows how biases are interrelated and reflect broader practices of discrimination and exclusion. Regarding AI safety, contextual dynamics shape the narrative of what constitutes an "existential threat" (Gebru and Torres 2024), often certain hegemonic values and depictions of threat over others.
- **Combat neutrality:** A focus on neutrality examines how AI policy acts to promote a universal notion of impact, harm and threat. As the experiences of AI for those "at the margins" (Collins 1986) are different than those with substantial levels of privilege, a presumption of neutrality fails to represent these diverse experiences. When safe AI is designed, it often is encoded with a presumed universal benevolence. This is evident in the "character training" involved in Claude 3.0, where developers made a list of traits they wanted the model to have, leading it to generate and rank responses in accordance with these traits (Anthropic 2024; Bai et al. 2022). While this approach identifies how these systems are trained, there are important considerations that impact what traits are valued and considered universal.
- **Increase power for those at the margins:** Investigating power involves questioning not only who has a seat at the table, but also what that seat looks like. Rather than seeing governance as a top-down technocratic process, an intersectional feminist policy framework argues that there is a need for meaningful participation from those who are disproportionately marginalized by an AI system. Where safety movements may push for overarching technical solutions, this perspective sees solution building as a holistic process encompassing policy, design, literacy and justice.

Figure 1: A Feminist AI Policy Framework



Source: Author.

The feminist AI policy framework was used to examine seven international initiatives regarding AI safety governance. These documents were chosen for their prominence in the AI safety movement, as well as their diverse scope and varying approaches: state-led, industry-led or overseen by an international body (see Figure 2). The following section details how the feminist AI policy framework was used to understand the ways in which the aforementioned goals were or were not achieved.

Figure 2: AI Safety Initiatives

	INTIATIVE	SIGNATORIES/INITATOR
INT.ORG	Recommendation of the Council on OECD Legal Instruments Artificial Intelligence (OECD Safe AI Principles)	OECD
	Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development	UN General Assembly, initiated by USA
STATE	Seoul Declaration for Safe, Innovative and Inclusive AI	Australia, Canada, EU, France, Germany, Italy, Japan, Korea, Singapore, UK, USA
	Bletchley Declaration	Australia, Brazil, Canada, Chile, China, EU, France, Germany, India, Indonesia, Ireland, Israel, Italy, Japan, Kenya, Netherlands, Nigeria, The Philippines, Korea, Rwanda, Saudi Arabia, Singapore, Spain, Switzerland, Türkiye, Ukraine, UAE, UK, USA
INDUSTRY	Frontier AI Safety Commitments	Amazon, Anthropic, Cohere, Google, G42, IBM, Inflection AI, Meta, Microsoft, Mistral AI, Naver, OpenAI, Samsung Electronics, Technology Innovation Institute, xAI, Zhipu.ai
	Asilomar AI Principles	Coordinated by Future of Life Institute

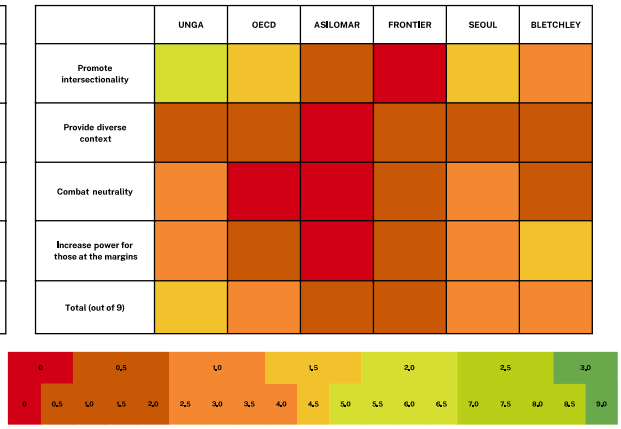
Source: Author.

Findings

Utilizing the framework, this research found that none of the seven documents successfully meet each of the goals. Figure 3 details the corresponding scores of each initiative. For each goal, the initiatives could receive up to three points — one per question, for a maximum framework score of nine. The right side of Figure 3 indicates a heat map of the scores, with the lowest (zero) being represented with red, and the highest possible score (three for individual goals, nine for framework total) being represented with green.

Figure 3: Findings from Analysis

	UNGA	OECD	ASILOMAR	FRONTIER	SEOUL	BLETCHLEY
Promote intersectionality	2.0	1.5	0.5	0	1.5	1.0
Provide diverse context	0.5	0.5	0	0.5	0.5	0.5
Combat neutrality	1.0	0	0	0.5	1.0	0.5
Increase power for those at the margins	1.0	0.5	0	0.5	1.0	1.5
Total (out of 9)	4.5	2.5	0.5	1.5	4	3



Source: Author.

With regards to *promote intersectionality*, the scores demonstrate how there is a varied commitment to examining the intersecting identities that are facing existential threats. While no initiative achieved a perfect score, the UN resolution titled "Seizing the opportunities of safe, secure, and trustworthy artificial intelligence systems for sustainable development" calls upon members to close the gender digital divide, encouraging them to "mainstream a disability, gender and racial equality perspective in policy decisions."[2] This contrasts with the lower scoring approaches, which did not directly identify the need for intersectional perspectives. For example, the Asilomar Principles state that: "The goal of AI research should be to create not undirected intelligence, but beneficial intelligence."[3] This approach highlights the need for beneficial AI but does not address the need to examine what constitutes "beneficial AI" for groups who are actively experiencing disproportionate harms.

When examining how initiatives *provide diverse context*, the scores exhibited a smaller range, with the majority of documents scoring 0.5. These low scores reflect the AI safety movement's insufficiency in connecting future threats to current existential harms being faced by marginalized groups. For example, the Organisation for Economic Co-operation and Development (OECD) details a very limited definition of AI knowledge: "the skills and resources, such as data, code, algorithms, models, research, know-how, training

2 *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*, GA Res 78/265, UNGAOR, 78th Sess, UN Doc A/78/L.49 (2024) at 6(p).

3 See https://futureoflife.org/open-letter/ai-principles/.

programmes, governance, processes, and best practices required to understand and participate in the AI system lifecycle, including managing risks."[4] This definition fails to incorporate lived experiences and diverse knowledge into the development of AI safety plans. Additionally, no initiative examined the biases associated with AI safety research, including the overrepresentation of white males, and the reinforcement of narratives surrounding technological utopianism, inevitable progress and effective altruism (Gebru and Torres 2024).

Similar conclusions were found regarding efforts to *combat neutrality*. The language surrounding existential risks often depicted threats as both future and universal. For example, numbers 14 and 15 of the Asilomar Principles detail a commitment to shared AI benefits and prosperity. However, they fail to detail how these shares will be distributed according to equitable need (substantive equality) as opposed to provision of a single level of support and resources for all (formal equality). When describing AI risk, the Seoul Declaration states, "Recognizing that all states will be affected by the benefits and risks of AI, we will actively include a wide range of international stakeholders in conversations around AI governance" (GOV.UK 2024). This presumed universal risk ignores how certain groups experience threats that could be considered as existential, such as a discriminatory algorithm that causes a low-income family to lose their social welfare (Eubanks 2018) or a data centre with excessive cooling regimens that lead to a lack of accessible water for rural and Indigenous communities (Valdivia 2024). When a policy seeks to only address future threats, it fails to understand how current harms pose existential threats to vulnerable communities.

When addressing how to *increase power for those at the margins*, these initiatives displayed low scores. To achieve this goal, initiatives needed to have a clear plan with regards to ensuring diverse participation. The Asilomar AI Principles[5] and Frontier AI Safety Commitments[6] appear to be mainly symbolic, reiterating Gebru and Torres' (2024) perspective that corporate actors utilize AI safety as a way to maintain control over AI development. In developing vague ethical guidelines or voluntary commitments, corporate actors are able to deflect from social pressures for strong regulation while affirming their role in the governance process. Furthermore, these initiatives display a focus on expert consultation, without considering how equitable AI requires diverse forms of knowledge (Stinson and Vlaad 2024). In particular, the OECD Safe AI Principles highlight a "network of experts," the majority of whom represent government, academia and industry. There is no substantial participation of organizations or civil society groups specializing in marginalized perspectives or alternative epistemologies, reflecting the OECD's restrictive definition of what constitutes AI knowledge. Similarly, the Bletchley Declaration outlines the creation of an inclusive network of scientific research on frontier AI safety.[7] The declaration outlines multilateral, bilateral and plurilateral collaboration, but fails to acknowledge how they plan to meaningfully cultivate "inclusivity" in this network.

4 OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (2023), online: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

5 See https://futureoflife.org/open-letter/ai-principles/.

6 See www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.

7 See GOV.UK (2023).

Lessons Learned: The Future of Feminist AI Safety Research

In examining these initiatives through the feminist AI policy framework, it is clear that the AI safety movement requires greater engagement with feminist protocols. This concluding section builds upon these findings to identify two key values that can be used proactively to enact feminist AI safety practices: accountability and participation. In addition to these values, policy makers are encouraged to utilize the framework outlined in Figure 1 to conduct a rigorous feminist analysis of proposed policies.

Accountability: Future Risks versus Current Harms

Biased AI technology has significant impacts on marginalized groups, including women (Crawford 2021), racialized people (Schelenz 2022; Benjamin 2019) and economically subjugated groups (Nopper 2020), representing greater hierarchies of oppression and power. As Safiyah Umoja Noble (2018) highlights, this pattern leads to "technological redlining" where certain groups are systematically disadvantaged.

Within the AI safety movement, there is little mention of these current harms. The embrace of AI safety as a utopian goal has enabled corporate actors to evade responsibility for these current AI harms (Gebru and Torres 2024). Current understandings of both technical knowledge (Stinson and Vlaad 2024) and who is responsible for constructing ethical AI (Browne, Drage and McInerney 2024) highlight the need to develop stronger accountability measures. Tom Slee (2020) argues that relying on private sector soft law will not ensure accountability, as the guiding values of the tech industry, such as profit maximization, are incompatible with the promotion of regulatory safeguards. As highlighted by Sophie Toupin (2024), a feminist AI policy should be constructed as something that does not produce lofty and symbolic goals but recognizes the interconnected practices of discrimination, bias, extraction and exploitation that constitute AI. In scrutinizing the AI safety movement's focus on universalism and soft law, there can be a move toward meaningful accountability structures.

Participation: Existential for Whom?

At the AI Safety Summit in 2023, then-US Vice President Kamala Harris noted: "threats are often referred to as the 'existential threats of AI' because, of course, they could endanger the very existence of humanity. These threats, without question, are profound, and they demand global action. But let us be clear. There are additional threats that also demand our action — threats that are currently causing harm and which, to many people, also feel existential" (Harris 2023).

Harris then listed potential threats, which also feel existential to certain groups, including abuse from the threat of explicit deep-fake photographs, wrongful imprisonment and AI-enabled misinformation. Amba Kak, the executive director of AI Now, further argues that AI safety ought to be designed and implemented by those who are most impacted (Kak 2023). Figure 4 details how systemic biases create unjust impacts that become normalized when they are collected as seemingly objective data (D'Ignazio and Klein 2020; Baack 2024). When this data is then used to train AI models, the models reify these biases as normal and objective (Gillespie 2024).

Figure 4: Breaking Down AI Existential Risk

Source: Author.

If these impacts continue to go unaddressed at the systemic level, with solutions prioritizing transparency (Ananny and Crawford 2018), anti-discrimination discourse (Hoffmann 2019) and corporate control (Stark, Greene and Hoffmann 2020) then the development of AGI will be imbued with these systemic harms. AI safety needs to embrace multifaceted, multi-stakeholder participation to address situated existential risk, as those at the margins often feel existential risk first (Lucero-Matteucci 2023). While there is a definite barrier to ensuring meaningful participation, a commitment to developing feminist AI involves not only a commitment to non-discrimination, but also a commitment to increased digital literacy, investment into public intervention strategies such as citizen juries or mini publics (Brandusescu and Sieber 2023) and available grants to develop community-focused safe AI. A proactive approach would enable pathways for meaningful participation and inclusion of various stakeholders through pre-emptive input from marginalized communities and shifting the power to use and deploy algorithms to the communities in which they will be utilized (Okidegbe 2022).

Much of the discussions surrounding AI safety present similar concerns to those argued by feminist AI scholars — both viewpoints are concerned with the potential future of AI, both want to establish clear safeguards and both challenge the idea that AI is universally beneficial. However, within the AI safety movement, there is a clear lack of diverse perspectives. AGI will reflect the biases we are already seeing in AI today. To address gaps in AI safety, there is a need to highlight how certain groups are experiencing forms of existential risks due to the detrimental harms posed by AI. Policy design must acknowledge the interconnectedness of systemic biases and future AI risk. In merging AI safety with feminist theory, this paper has argued for an account that challenges current power structures to prevent future risks.

Recommendations

- **Recommendation 1:** Policy makers should conduct rigorous audits of AI safety policies utilizing the Feminist AI Policy Framework both pre- and post hoc, to ensure an ongoing commitment to intersectionality.
- **Recommendation 2:** AI safety initiatives should incorporate attempts to address risks that are existential at differing intersectional levels by holding corporate actors accountable for discrimination, extraction and exploitation occurring from their AI systems.
- **Recommendation 3:** Future AI safety policies should underscore that future existential threats are grounded in current harms and need to be addressed using a diverse set of technical knowledge *and* lived experience throughout the policy life cycle.

Acknowledgements

I am incredibly grateful to my various peer reviewers, including my colleague Nathalie DiBerardino, and my hub mentor Maroussia Lévesque, who both provided thorough and detailed suggestions. I am also thankful to have the support of Reanne Cayenne and Dianna English at CIGI, who have been strong guiding lights throughout my Hub fellowship. Additionally, I am grateful to Paul Moore at Toronto Metropolitan University and Mitacs for the generous support in pursuing this work.

About the Author

Laine McCrory is a Digital Policy Hub master's fellow and second-year master's student in the joint program in communication and culture at Toronto Metropolitan University and York University. She works at the intersections of feminist technology, artificial intelligence (AI) policy, smart cities, data capture and community governance in order to create socio-political critiques of AI.

Works Cited

Acemoglu, Daron. 2021. "The AI we should fear is already here." *The Washington Post*, July 21. www.washingtonpost.com/opinions/2021/07/21/ai-we-should-fear-is-already-here/.

Ananny, Mike and Kate Crawford. 2018. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." *New Media & Society* 20 (3): 973–89. https://doi.org/10.1177/1461444816676645.

Anthropic. 2024. "Claude's Character." Anthropic, June 8. www.anthropic.com/research/claude-character.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." *arXiv*, July 25. https://doi.org/10.48550/arxiv.1606.06565.

Baack, Stephen. 2024. “Training Data for the Price of a Sandwich: Common Crawl’s Impact on Generative AI.” Mozilla Foundation, February 6. https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen et al. 2022. “Constitutional AI: Harmlessness from AI Feedback.” *arXiv*, December 15. https://doi.org/10.48550/arxiv.2212.08073.

Bender, Emily M., Timrit Gebru, Angelina McMillan-Major and Shmargret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” *FAccT’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. https://doi.org/10.1145/3442188.3445922.

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

Bensimon, Estela Mara and Catherine Marshall. 2003. “Like It or Not: Feminist Critical Policy Analysis Matters.” *The Journal of Higher Education* 74 (3): 337–49. https://doi.org/10.1080/00221546.2003.11780850.

Bostrom, Nick. 2002. “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards.” *Journal of Evolution and Technology* 9. https://nickbostrom.com/existential/risks.pdf.

– – –. 2014. *Superintelligence: Paths, Dangers, Strategies*. 1st ed. Oxford, UK: Oxford University Press.

Brandusescu, Ana and Renee Sieber. 2023. “Canada’s Artificial Intelligence and Data Act: A missed opportunity for shared prosperity.” SSRN. http://dx.doi.org/10.2139/ssrn.4602943.

Browne, Jude, Eleanor Drage, and Kerry McInerney. 2024. “Tech workers’ perspectives on ethical issues in AI development: Foregrounding feminist approaches. *Big Data & Society* 11 (1). https://doi.org/10.1177/20539517231221780.

Buolamwini, Joy and Timrit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” *Proceedings of the 1st Conference on Fairness*, *Accountability and Transparency* 81: 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Cellan-Jones, Rory. 2014. “Stephen Hawking warns artificial intelligence could end mankind.” BBC News, December 2. www.bbc.com/news/technology-30290540.

Cohere Team. 2022. “A Joint Recommendation for Language Model Development.” *The Cohere Blog*, June 2. https://cohere.com/blog/best-practices-for-deploying-language-models.

Collins, Patricia Hill. 1986. “Learning from the Outsider Within: The Sociological Significance of Black Feminist Thought.” *Social Problems* 33 (6): s14–s32. https://doi.org/10.2307/800672.

Crawford, Kate. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. 1st ed. New Haven, CT: Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t.

Crenshaw, Kimberlé. 1991. “Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color.” *Stanford Law Review* 43 (6): 1241–99. https://doi.org/10.2307/1229039.

D’Ignazio, Catherine and Lauren F. Klein. 2020. *Data Feminism.* Cambridge, MA: MIT Press.

Eadicicco, Lisa. 2015. “Bill Gates: Elon Musk Is Right, We Should All Be Scared Of Artificial Intelligence Wiping Out Humanity.” *Business Insider*, January 28. www.businessinsider.com/bill-gates-artificial-intelligence-2015-1.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. London, UK: St. Martin’s Press.

Fung, Brian. 2023. “UN Secretary General embraces calls for a new UN agency on AI in the face of ‘potentially catastrophic and existential risks.’” CNN, July 18. www.cnn.com/2023/07/18/tech/un-ai-agency/index.html.

Future of Life Institute. 2023. “Pause Giant AI Experiments: An Open Letter.” March 22. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gebru, Timnit and Émile P. Torres. 2024. “The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence.” *First Monda*y 29 (4). https://doi.org/10.5210/fm.v29i4.13636.

Gillespie, Tarleton. 2024. “Generative AI and the politics of visibility.” *Big Data & Society* 11 (2). https://doi.org/10.1177/20539517241252131.

GOV.UK. 2023. “The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023.” www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

– – –. 2024. “Seoul Declaration for safe, innovative and inclusive AI by participants attending the Leaders’ Session: AI Seoul Summit, 21 May 2024.” Department for Science, Innovation & Technology. www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-declaration-for-safe-innovative-and-inclusive-ai-by-participants-attending-the-leaders-session-ai-seoul-summit-21-may-2024.

Hamidi, Foad, Morgan Klaus Scheuerman and Stacy M. Branham. 2018. “Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems.” In *CHI’18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173582.

Hankivsky, Olena and Renee Cormier. 2011. “Intersectionality and Public Policy: Some Lessons from Existing Models.” *Political Research Quarterly* 64 (1): 217–29. https://doi.org/10.1177/1065912910376385

Hankivsky, Olena, Daniel Grace, Gemma Hunting, Melissa Giesbrecht, Alycia Fridkin, Sarah Rudrum, Olivier Ferlatte and Natalie Clark. 2014. “An intersectionality-based policy analysis framework: critical reflections on a methodology for advancing equity.” *International Journal for Equity in Health* 13 (1). https://doi.org/10.1186/s12939-014-0119-x.

Harris, Kamala. 2023. “Remarks by Vice President Harris on the Future of Artificial Intelligence, London, United Kingdom.” AI Safety Summit speech, November 1. https://bidenwhitehouse.archives.gov/briefing-room/speeches-remarks/2023/11/01/remarks-by-vice-president-harris-on-the-future-of-artificial-intelligence-london-united-kingdom/.

Hendrycks, Dan, Mantas Mazeika and Thomas Woodside. 2023. “An Overview of Catastrophic AI Risks.” *arXiv*, October 9. https://doi.org/10.48550/arxiv.2306.12001.

Hoffmann, Anna Lauren. 2019. “Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse.” *Information, Communication & Society* 22 (7): 900–15. https://doi.org/10.1080/1369118X.2019.1573912.

Hyde, Cheryl. 2000. “Feminist Approaches to Social Policy.” In *The Handbook of Social Policy*, edited by James Midgley, Martin B. Tracy and Michelle Livermore, 421–35. Thousand Oaks, CA: Sage.

Kak, Amba. 2023. “Remarks from AI Now ED Amba Kak on Day 2 of the UK AI Safety Summit.” Speech at AI Safety Summit, November 2. https://ainowinstitute.org/news/events/remarks-from-ai-now-ed-amba-kak-on-day-2-of-the-uk-ai-safety-summit.

Kanenberg, Heather. 2013. “Feminist Policy Analysis: Expanding Traditional Social Work Methods.” *Journal of Teaching in Social Work* 33 (2): 129–42. https://doi.org/10.1080/08841233.2013.772935.

Kanenberg, Heather, Roberta Leal and Stephen “Arch” Erich. 2020. “Revising McPhail’s Feminist Policy Analysis Framework: Updates for Use in Contemporary Social Policy Research.” *Advances in Social Work* 19 (1): 1–22. https://doi.org/10.18060/22639.

Klein, Lauren and Catherine D’Ignazio. 2024. “Data Feminism for AI.” In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 100–12. New York, NY: ACM. https://doi.org/10.1145/3630106.3658543.

Lazar, Seth and Alondra Nelson. 2023. “AI safety on whose terms?” *Science* 381 (6654): 138. https://doi.org/10.1126/science.adi8982.

Lucero-Matteucci, Kayla. 2023. “Catastrophic risks are converging. It’s time for researchers to step out of their silos.” Bulletin of the Atomic Scientists, May 1. https://thebulletin.org/2023/05/catastrophic-risks-are-converging-its-time-for-researchers-to-step-out-of-their-silos/.

Mansfield, Katherine Cumings, Anjalé D. Welton and Margaret Grogan. 2014. “‘Truth or consequences’: a feminist critical policy analysis of the STEM crisis.” *International Journal of Qualitative Studies in Education* 27 (9): 1155–82. https://doi.org/10.1080/09518398.2014.916006.

McPhail, Beverly A. 2003. “A Feminist Policy Analysis Framework: Through a Gendered Lens.” *Social Policy Journal* 2 (2–3): 39–61. https://doi.org/10.1300/J185v02n02_04.

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

Nopper, Tamara K. 2020. “Digital Character in ‘The Scored Society’ FICO, Social Networks, and Competing Measurements of Creditworthiness.” In *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*, edited by Ruha Benjamin, 170–87. Durham, NC: Duke University Press.

Okidegbe, Ngozi. 2022. “The Democratizing Potential Of Algorithms?” *Connecticut Law Review* 53 (739). https://scholarship.law.bu.edu/faculty_scholarship/3138.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. New York, NY: Hachette.

Ricaurte, Paola. 2024. “How can feminism inform AI governance in practice?” UNESCO, February 1. www.unesco.org/en/articles/how-can-feminism-inform-ai-governance-practice.

Roche, Cathy, P. J. Wall and Dave Lewis. 2022. “Ethics and diversity in artificial intelligence policies, strategies and initiatives.” *AI and Ethics* 3 (4): 1095–115. https://doi.org/10.1007/s43681-022-00218-9.

Schelenz, Laura. 2022. “Artificial Intelligence Between Oppression and Resistance: Black Feminist Perspectives on Emerging Technologies.” In *Artificial Intelligence and Its Discontents*, edited by Ariane Hanemaayer, 225–49. Cham, Switzerland: Springer International. https://doi.org/10.1007/978-3-030-88615-8_11.

Slee, Tom. 2020. “The Incompatible Incentives of Private-Sector AI.” In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale and Sunit Das, 107–23. New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.6.

Sotala, Kaj and Roman V. Yampolskiy. 2014. “Responses to catastrophic AGI risk: a survey.” *Physica Scripta* 90: (1): 018001. https://doi.org/10.1088/0031-8949/90/1/018001.

Stark, Luke, Daniel Greene and Anna Lauren Hoffmann. 2020. “Critical Perspectives on Governance Mechanisms for AI/ML Systems.” In *The Cultural Life of Machine Learning*, edited by Jonathan Roberge and Michael Castelle, 257–80. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-56286-1_9.

Stinson, Catherine and Sofie Vlaad. 2024. “A feeling for the algorithm: Diversity, expertise, and artificial intelligence.” *Big Data & Society* 11 (1). https://doi.org/10.1177/20539517231224247.

Tacheva, Jasmina and Srividya Ramasubramanian. 2023. “AI Empire: Unraveling the interlocking systems of oppression in generative AI’s global order.” *Big Data & Society* 10 (2). https://doi.org/10.1177/20539517231219241.

Toupin, Sophie. 2024. “Shaping feminist artificial intelligence.” *New Media & Society* 26 (1): 580–95. https://doi.org/10.1177/14614448221150776.

Ulnicane, Inga and Aini Aden. 2023. “Power and politics in framing bias in Artificial Intelligence policy.” *Review of Policy Research* 40 (5): 665–87. https://doi.org/10.1111/ropr.12567.

Valdivia, Ana. 2024. “The supply chain capitalism of AI: a call to (re)think algorithmic harms and resistance through environmental lens.” *Information, Communication & Society*, 1–17. https://doi.org/10.1080/1369118X.2024.2420021.

Variengien, Alexandre and Charles Martinet. 2024. “AI Safety Institutes: Can countries meet the challenge?” OECD AI Policy Observatory, July 29. https://oecd.ai/en/wonk/ai-safety-institutes-challenge.

Vold, Katrina and Daniel R. Harris. 2021. “How Does Artificial Intelligence Pose an Existential Risk?” In *The Oxford Handbook of Digital Ethics*, edited by Carissa Véliz, 724–47. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.36.

Voss, Peter. 2017. “What is AGI?” Medium — Intuition Machine, February 21. https://medium.com/intuitionmachine/what-is-agi-99cdb671c88e.