Feminist AI: Confronting Bias, Power, and the Promise of Justice-Oriented Technology

Rashmi Singh
Dept. of Political Science
Delhi University, New Delhi
Email: raasshi9026@gmail.com

Abstract

As AI becomes increasingly integrated into critical social, economic, and political institutions — from healthcare delivery to criminal justice administration to employment screening to social media governance — urgent questions arise about whose interests these technologies serve and whose epistemological frameworks are sidelined. In this paper, we map the path of AI development through an intersectional feminist lens to argue that the future of artificial intelligence cannot be a function only of progress in technical efficiency but must reckon with systems of structural power long woven into the fabric determining which knowledge is pursued and who has access to it during research, development and deployment.

Although artificial intelligence offers us new possibilities for social progress—improved diagnostic materials, accessibility technologies, aids to liberatory movements such as #MeToo and Black Lives Matter—it also poses profound challenges in its current configuration. Algorithmic bias causes systematic discrimination against women, racialized communities, and other marginalized groups — facial recognition technologies have been shown to have differential accuracy rates across demographic categories; hiring algorithms have replicated occupational segregation; and predictive policing systems replicate historical patterns of over-policing of minority communities. The emergence of deepfake technology has added a novel tool to the repertoire of technology-facilitated gender-based violence, engaged under regulatory lacunae disproportionately against women in the form of non-consensual intimate image abuse. Moreover, the environmental extractivism and exploitation of Global South data laborers incorporated in large-scale AI models reflect approaches to colonial extraction central to contemporary forms of technological innovation.

Keywords: *data feminisms, algorithmic injustice, Homo digitalis, techno-solutionism, artificial intelligence, deepfakes*

Reference to this paper should be made as follows:
Received: 25.03.26
Approved: 12.06.26

Rashmi Singh

Feminist AI: Confronting Bias, Power, and the Promise of Justice-Oriented Technology

Vol. XVII, No.1
Article No. 29,
Pg. 280-291

Similarity Check: 05%

Online available at
https://anubooks.com/journal-volume/jgv-vol-xvii-no1-jan-june-2026

DOI: https://doi.org/10.31995/jgv.2026.v17i01.029

**This article has been peer-reviewed by the Review Committee of JGV.*

https://doi.org/10.31995/jgv.2026.v17i01.029

Introduction

The brisk spread of artificial intelligence into all corners of social life has brought with it a parallel boom in scholarly and public concern about the ethical, social, and political consequences of these technologies. From predictive policing systems that drive amplified racial profiling (O'Neil, 2016) to hiring algorithms with the potential to exacerbate occupational segregation (Raghavan et al., 2020), and from large language models that are known to replicate gender stereotypes (Kotek et al., 2023) through generative tools enabling non consensual intimate imagery (Akter & Ahmed, 2024), the tropes of AI's ability to seal inequality have emerged at a disconcertingly fast speed. Yet amidst increased recognition of these harms, the predominant response from industry and much of the technical community is still tethered to what critics describe as "techno-solutionism" — the idea that social ills can be solved through better algorithms, more data, or other forms of incremental technical fix (Satra & Selinger, 2023; Barnes, 2022).

Within this context, feminist scholarship provides an indispensible corrective. Feminism's analytic mode of attuning to structural power, its insistence on the (practical and often material) interlocking nature of oppression, and its commitment to situated, embodied knowledge provide resources for understanding not just that AI systems are harmful but also why they are harmful and how those harms may be addressed at their root (D'Ignazio & Klein, 2020; Klein & D'Ignazio, 2024).). And intersectional feminism, in particular, sheds light on how algorithmic discrimination overlaps and multiplies across gender, race, class and geography - a complexity that is frequently erased in narrowly technical framings of "bias" (Crenshaw, 1989; Hampton, 2021).

Feminist approaches have started to become more visible within the FAccT (Fairness, Accountability, and Transparency) community. Hampton (2021) promoted Black feminism as a way of seeing algorithmic oppression, which highlights how race and gender together create special harms. Hancock Li and Kumar (2021) used a feminist epistemology lens to interrogate feature importance methods; Klumbyte, Draude, and Taylor (2022) also showed how ideas from feminist science and technology studies—such as agential realism or situated knowledges—may be operationalized in machine learning design. Outside of conference proceedings, twenty one chapters appearing in the edited volume Feminist AI (Browne et al., 2024) across a range of topics on Afro feminist digital futures, the political economy of AI and feminist approaches to governing AI.

Literature Review

1. Origins of Feminist AI Research

But they join an intellectual tradition of feminist engagement with artificial intelligence that goes back at least to Alison Adam's (1998) Artificial Knowing, which used insights from feminist epistemology to undermine universalizing claims in early AI research. In Adam's view the AI systems that so often pride themselves on being neutral, objective instruments of rationality are actually less independent entities than they are situated avatars for their mostly male Western developers—a critique that jibes strikingly with current anxieties around biases in training data and the homogenous demographics of corporate AI research groups (Wajcman & Young, 2023). More recently, Keyes and Creel (2022) have reprised Adam's perceptions by linking them to a vision of feminist AI focused on "whose knowledge and interests be "reflected" in algorithmic systems.

Feminist approaches have started to become more visible within the FAccT (Fairness, Accountability, and Transparency) community. Hampton (2021) promoted Black feminism as a way of seeing algorithmic oppression, which highlights how race and gender together create special harms. Hancock Li and Kumar (2021) used a feminist epistemology lens to interrogate feature importance methods; Klumbyte, Draude, and Taylor (2022) also showed how ideas from feminist science and technology studies—such as agential realism or situated knowledges—may be operationalized in machine learning design. Outside of conference proceedings, twenty one chapters appearing in the edited volume Feminist AI (Browne et al., 2024) across a range of topics on Afrofeminist digital futures, the political economy of AI and feminist approaches to governing AI.

2. Algorithmic Bias and Intersectionality

A vast literature documents the magnitude and scope of algorithmic bias across various domains — employment, housing, criminal justice and health care. Noble's (2018) Algorithms of Oppression showed how search engines perpetuate racist and sexist stereotypes, while O'Neil's (2016) Weapons of Math Destruction explained how risk assessment tools disproportionately disadvantage marginalized communities. More recently, Buolamwini and Gebru's (2018) study "Gender Shades" showed that commercial facial recognition systems have higher error rates for darker skinned women: an intersectional disparity that cannot be accounted for by gender or race in isolation.

3. Deepfakes and Technology Facilitated Gender Based Violence

A newer but fast-growing strand of literature explores the use of AI generated deepfakes as a weapon for gender based violence. Deepfake pornography accounts for around 90 per cent of deepfake content online and targets women overwhelmingly (Akter & Ahmed, 2024) according to a systematic overview. This type of image based sexual abuse causes intense psychological harm, destroys professional and personal reputations, and is used to silence or blackmail victims (see Citron & Chesney, 2019; Henry et al., 2020).

Feminist scholars have positioned deepfake abuse as part of longer histories of gendered violence mediated through technology. According to Eaton and McGlynn (2020), deepfakes show a new way of asserting power and control, while Van Der Nagel (2020) points out that they "take on, propagate and bolster an online space in which images of women are manufactured as products for men's consumption." Yet despite these harms, regulatory responses are still fragmented (Hao, 2021) and content moderation systems often continue to be unable to detect or remove deepfake content in a timely manner. Given the disproportionate impact on women, especially of public figures and activists, it highlights a need for feminist analysis and intervention.

4. Political Economy of AI : Labor, Environment and Colonial Dynamics

A fourth body of literature grounds AI within wider structures of racial capitalism and colonialism. Academics have recorded the extractive labor processes upon which AI is built: data annotation and content moderation are frequently farmed out to Global South workers who are paid poverty wages and subject to exposure of traumatic material, (Posada, 2022; Perrigo, 2023). Again, this division of labor reproduces colonial patterns whereby high value intellectual work occurs in the Global North and precarious low paid labor is transferred to the Global South — something that Couldry and Mejias (2019) call "data colonialism."

At the same time, the environmental impact of AI has come under critical scrutiny. Once trained, many of these models are fine-tuned for specific tasks and yield impressive results that often become state-of-the-art (Li et al., 2023). Data centers tend to be sited in communities with lower political power, and climate change's effects — worsened the energy demands from A.I. — disproportionately impact populations in the Global South

Literature Gap

Although the literature reviewed above demonstrates a body of rich scholarship on feminist critiques of AI, several gaps remain. Second, most existing work analyzes isolated harms (e.g., bias in a given system) rather than interlocking structural dynamics that generate those harms across domains. Second, while frequently invoked, empirical studies that collectively consider and operationalize intersectional analysis in the development and evaluation of AI products are scarce (Kong 2022). Third, the connection between AI facilitated gender-based violence on one hand (e.g., deepfakes), and other forms of harm AI might cause (e.g., labor exploitation, environmental degradation) is rarely analyzed concurrently, which leads to piecemeal responses. Fourth, while feminist critiques of techno-solutionism are well-articulated, there remains fewer studies that have offered tangible, actionable frameworks for structural transformation capable of marrying technical, regulatory and social movement approaches.

Methodology

1. Research Design

As such, this study draws on a qualitative, interpretive research design grounded in feminist epistemology. '! Feminist methodology: Reflexivity, focus on power dynamics in research and the importance of situated, embodied knowledge (Harding, 1987) Due to the conceptual and integrative nature of the inquiry, this study does not gather primary empirical data, but rather weaves existing scholarship together in developing an analytical framework.

2. Data Sources

Bibliographic data comes from academic literatures in feminist theory, science and technology studies, computer science (where the focus is on FAccT and related venues), media studies, and legal studies that have been peer reviewed. The review further includes gray literature from civil society (e.g., Coding Rights, IT for Change) and prescriptive journalism documenting AI harms in contexts usually missing from the academic literature. Articles were searched for in Google Scholar, Scopus, ACM Digital Library using keywords "feminist AI," "algorithmic bias," "deepfakes gender-based violence," data feminism', 'AI labor' and 'AI environment'.

3. Analytical Approach

The analysis is carried out in three parts. First, we thematically review the literature, organized according to the four challenge areas highlighted in this literature review. Second, we apply the analytic lenses of intersectional feminism and data feminism to isolate shared structural patterns across these spaces. Third, this study synthesizes findings into a framework for structural intervention, utilizing thematic analysis to extract transailing themes and the proposed structure of interventions.

4. Reflexivity and Positionality

Following feminist research practices, they declare their positionality: as an academic-based researcher in the Global North with resources and platforms unevenly available worldwide. The analysis seeks to foreground perspectives from and the knowledge produced by marginalized communities through emphasizing scholarship by women of color, Indigenous feminists, and Global South researchers; and paying attention to sites of knowledge production outside the academy, such as activism, art, and community organizing.

Analysis / Findings

The literature synthesis yielded four overarching themes, each representing a challenge domain. These results are presented in Table 1 and conceptually illustrated in Figure 1.

1. Theme 1: Algorithmic Bias is Structural Rather than Incidental

Algorithmic bias appeared in all studies fielded, and it was more a predictable consequence of asymmetries in structural power than an accidental "glitch." Bias in the data is enshrined through three mechanisms; (a) training data that mirror historical discrimination, (b) features selected that privilege the norms of dominant groups, and (c) evaluation metrics that invisibilize intersectional harm. Specifically, fairness interventions that optimize for demographic parity can miss subgroup disparities, illustrated in Figure 1, which shows how error rates change across gender-race combinations.

Table 1: Findings by Challenge Domain — Summary of Key Takeaways

Domain	Primary Mechanism of Harm	Feminist Critique	Proposed Intervention
Algorithmic bias	Historical discrimination reflected in data; narrow evaluation metrics	Single-axis fairness erases intersectional disparities; structural causes ignored due to focus on tech changes	Intersectional audits; participatory dataset curation; regulatory standards
Deepfakes & GBV	Non-consensual intimate image generation; weaponization of AI	Deepfakes as extension of patriarchal control; regulatory lacunae	Criminalization of non-consensual deepfakes; platform accountability; victim support
Extracting from labor	Outsourcing annotation/content moderation to Global South	Data colonialism; gendered/racialized division of labor	Worker organizing; living wage mandates; transparency in supply chains
Environmental Impact	Significant energy/water consumption, uneven distribution of impacts	How an ecofeminist lens connects environmental and gender justice	Sustainable AI research; public-interest infrastructure; reparative policies.

2. Theme 2: Deepfakes as a Gendered Weapon

The investigation found that the use of deepfake technology overwhelmingly focuses on committing gender-based violence. While estimates of the nature of such deepfake content vary between sources, decrees and reports from cybersecurity experts have estimated that roughly 96% percent of deepfake videos on the web are pornographic in nature, and a staggering 99 percent of each video was done not only as such but targeted women (see figure 2). The harms go beyond individual victims, producing a chilling effect on women's public participation — especially for activists, journalists and politicians. Scholars of feminism said this is a continuum of patriarchal violence, now enhanced through the use of AI.

3. Theme 3: AI Labor as Colonial Extraction

What the literature showed was that the AI industry is built on a world-wide supply chain of precari-ous labor. In contrast however, workers in Kenya, the Philippines and India do content moderation and data labeling work for wages well below living standards (Perrigo, 2023; Posada, 2022) often suffering psychological trauma without proper support. This division of labor perpetuates colonial power dynamics whereby the Global North captures value and externalizes costs. Feminist labor scholars pointed out the racialized and gendered composition of this invisible workforce: most data annotation workers are women of color in the Global South.

4. Theme 4: Environmental Harms Are Unequally Distributed

The analysis of AI's environmental footprint showed striking inequities. The carbon released for training a single large language model can amount to that of five cars over their total expected miles driven, and data centers use up water in areas already facing scarcity (Li et al., 2023). These burdens are disproportionately shouldered by low income communities and communities of color. Ecofeminist frameworks helped to mine the response, helping us elucidate how these patterns were continuous with the logic of extraction that also shapes relations of labor and algorithmic bias.

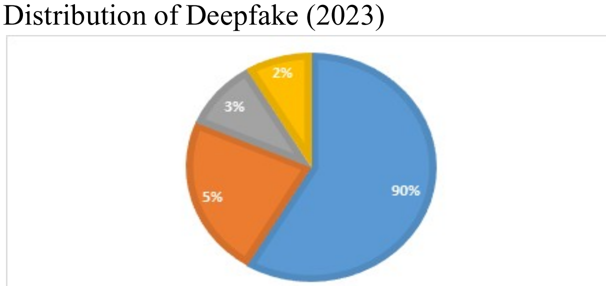
Figure 1

Commercial Facial Recognition: Error Rates by Intersectional Recurrentes (adapted from Buolamwini & Gebru, 2018)

Gender	Lighter skin	Darker skin
Male	0.8% error	12.0% error
Female	6.5% error	35.0% error

Figure 2

Content Type Distribution of Deepfake (2023)



https://doi.org/10.31995/jgv.2026.v17i01.029

Non-consensual pornography: 90%

Political disinformation: 5%

Satire/art: 3%

Other: 2%

Note: Data is compiled from Hao (2021) and Sensity (2021)

Discussion

The results confirm that the harms of AI are not simply local technical glitches but the expression of chronic structural inequalities. Algorithmic bias, deepfakes, labor extraction and environmental degradation are all interconnected by a common logic: the aggregation of power and resources into the hands of a few at the expense — whether through erasure or exploitation — of marginalized communities. Feminist frameworks—such as the critical concept of intersectionality and approaches like data feminism—expose these linkages and catalyze intervention. There is much wisdom to be integrated from this analysis and one of the critical learnings here has been techno-solutionism fails. In all four domains, technical fixes (bias mitigation algorithms, deepfake detection tools, efficiency improvements etc.) were deemed necessary but not sufficient. Without a response to the political economy that motivates AI development — corporate capture, absence of democratic oversight, and the deskilling of some forms of labour and knowledge (Mezzadra, 2014)— technical interventions could go on to reproduce the inequalities they claim to solve (Sloane et al., 2022; Barnes, 2022).

The findings also illuminate the importance of intersectionality. When we looked at a single-axis approach to fairness or harm reduction, it always failed the people at the intersections of multiple marginalized identities. Bias mitigation strategies that consider only gender parity, for example, can inadvertently exacerbate racial disparities; anti-deepfake legislation crafted without consideration for the unique vulnerabilities concerning LGBTQ+ users could leave them at risk. A feminist perspective requires that we are designing interventions from day 1 with intersectional analysis embedded.

A second major theme is the call for structural rather than individual accountability. In most existing mechanisms, the burden of proof is on and harm mitigation on victims — whether that's women negatively impacted by deepfakes, workers who face exploitation or communities that suffer environmental degradation. Feminist scholarship advocates shifting responsibility to the institutions — corporations, governments and research community that benefit from AI's development. This fits with Young's (2011) social connection model of responsibility, which argues that all who participate in unjust structures are at least co-responsible and so should contribute to their transformation.

Lastly, the insights suggest that collective action and solidarity are essential. The most promising interventions identified in the literature turned out to be bottom up organizing instead of top down technical solutions, such as data workers unionizing, grassroots feminist activists co designing tools together and cross sector coalitions advocating for public interest AI. These movements are rooted in feminist principles of relationality, care, and mutual aid, and they serve as an antidote to the individualism and the competition fostered by capitalist tech culture.

Conclusion

In this paper, I have argued that a feminist lens on artificial intelligence is vital for exploring and minimizing the systemic harms created by existing AI systems. Drawing together scholarship across four challenge domains—algorithmic unfairness, deepfakes are gender based violence, labor extraction, environmental harm—the study has shown that these harms can be seen as interconnected and predictable forms of structural inequalities deriving from race, gender, class and geography. Feminist frameworks, in particular intersectionality and data feminism, offer analytic tools to reveal these linkages and to go beyond techno solutionist responses that see inequality as a bug to be fixed rather than a feature of the political economy.

This finding has important implications for research, policy, and practice. For researchers, they highlight the importance of cross-disciplinary cooperation that foregrounds marginalized points of view and responds to the complete lifecycle of AI systems — from data sourcing through deployment and disposition. For policymakers, they indicate that regulatory frameworks cannot simply remedy bias and privacy; they must also consider labor conditions, environmental sustainability, and the concentration of corporate power. For practitioners they demand participatory design processes that position affected communities as knowledge holders rather than (unimaginably) stakeholders to consult.

The future of artificial intelligence itself will depend not on the technology alone, but on the values and institutions and movements that produce it. Feminism—long a utopian imagining otherwise—is a critical perspective and generative vision: technologies that optimize for collective flourishing rather than profit; arrangements structured around care, consent and ecology rather than resource extraction; institutions founded on disbursing power rather than concentrating it. The choice we face is not whether AI will exist, but whose interests it will prioritize — and feminism offers a model for making sure that it serves the many, rather than the few.

References

1. Adam, A. (1998). Artificial intelligence: Gender and the thinking machine Routledge.
2. Patter, Ah., & Mukelle (Home) 2024 Techno-feminist perspective on AI generated deepfakes as a new mechanism for gender-based violence against women: A narrative review. International Journal of Cyber Criminology, 18(1): Pg. **1–17**.
3. Barnes, M. R. (2022). “Better AI” Propaganda in Online Radicalization Feminist Philosophy Quarterly, 8(3/4), article 6.
4. Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code. Polity.
5. Browne, J., Cave, S., Drage, E. & McInerney, K. (Eds.). (2024). Feminist AI: Critical approaches to algorithms, data and intelligent machines Oxford University Press.
6. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Pg. **77–91.**
7. Cabral, L. (2010). Feminismos diversos: El feminismo comunitario. ACSUR Las Segovias.
8. (2019). Citron, D. K., & Chesney, R. Deep fakes: An emerging threat to privacy, democracy and national security. California Law Review, 107(6), Pg. **1753–1820.**
9. Couldry, N. & Mejias, U. A., 2019. Data colonialism: Rethinking big data’s relation to the contemporary subject. Television & New Media, 20(4), Pg. **336–349.**
10. Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics Pg. **139–167**, 1989(1) Univ. Chicago Legal Forum.
11. D’Ignazio, C., & Klein, L. (2020). Data feminism. MIT Press.
12. Eaton, A. A., & McGlynn, C. (2020). Nonconsensual porn: What it is and what you can do about it Policy Insights from the Behavioral and Brain Sciences, 7(2), Pg. **190–197.**
13. Hampton, L. M. (2021). Reflections by a black feminist on algorithmic oppression. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 1.

14. Hancock Li, L., & Kumar, I. E. (2021). A Solution: Epistemic values in feature importance methods — Lessons from feminist epistemology In Proceeding of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Pg. **817–826**).
15. Hao, K. (2021, February 12). Deepfake porn is destroying women's lives. Now the law might finally outlaw it. MIT Technology Review. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/
16. Harding, S. (1987). Social science issues in feminism and methodology. Indiana University Press.
17. Henry, N., McGlynn, C., Flynn, A., Johnson, K., Powell, A. & Scott, A. J. (2020). Potential impacts of non consensual nude or sexual imagery. Routledge.
18. Keyes, O., & Creel, K. (2022). Artificial knowing otherwise. Feminist Philosophy Quarterly, 8(3/4), Pg. **1–25.**
19. Klein, L., & D'Ignazio, C. (2024). Data feminism for AI. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.
20. Klumbyte, G., Draude, C. & Taylor, A. S. (2022). At the core of our understanding is access to words. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency: Pg. **1528–1541.**
21. Kong, Y. (2022). Are "intersectionally fair" A.I. algorithms really fair for women of color? A philosophical analysis. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAT* '22), Pg. **485–494.**
22. Li, P., Yang, J., Islam, M. A., & Ren, S.(2023). Less "thirsty" AI: Discovering and tackling the hidden water footprint of AI models. arXiv preprint arXiv:2304.03271.
23. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.
24. O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.
25. Perrigo, B. (2023, January 18). Exclusive: OpenAI hired workers in Kenya for low pay to make ChatGPT less toxic Time. https://time.com/6247678/openai-chatgpt-kenya-workers/
26. Posada, J. (2022). The coloniality of data work : Power and inequality in outsourced data production for machine learning (Doctoral dissertation). University of Toronto.

https://doi.org/10.31995/jgv.2026.v17i01.029

27. Satra, H. S., & Selinger E. (2023). The siren song of technological fixes for social issues: Defining, demarcating, and assessing techno fixes and techno solutionism. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4576801
28. Sloane, M., Moss, E., Awomolo, O. & Forlano, L. (2022). Participation is not the design solution to machine learning. Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Article No. 1.
29. Van Der Nagel, E. (2020). Deepfakes, control, and verification of images: Can we mark consent? Porn Studies, 7(4), Pg. **424–429.**
30. Wajcman, J. & Young, E. (2023). AI under scrutiny from feministic perspectives: Gender and digitalisation. In J. Browne, S. Cave, E. Drage, & K. McInerney (Eds.), Feminist AI: Critical perspectives on algorithms, data and intelligent machines Pg. **47–64.** Oxford University Press.
31. Young, I. M. (2011). Responsibility for justice. Oxford University Press.