Decoding Bias in Generative AI. Framing Socio-Technical Data Literacy as a Collective Critical Practice

Antonella Autuori[1,2]

[1] RMIT University, School of Design, Melbourne, Australia
[2] Institute of Design, SUPSI University of Applied Sciences and Arts of Southern Switzerland, Mendrisio, Switzerland

Abstract

This practice-based PhD develops a socio-technical data literacy framework for generative AI, emphasizing participatory engagement with the socio-cultural dimensions of data. Through methods such as participatory workshops and critical making, the research demonstrates how non-technical stakeholders can decode and intervene in algorithmic bias through engagement with classification data practices. The resulting toolkit and evaluative framework offer practical strategies for inclusive, culturally-aware participation in educational and civic contexts. By conceptualizing data literacy as a critical, situated and collective critical practice, the research contributes to HCI by advancing more equitable and relational human-AI interaction.

Keywords

generative artificial intelligence, bias, human-AI interaction, data literacy, feminist epistemologies

1. Introduction

Generative AI technologies are increasingly central to the infrastructures that mediate perception, representation, and decision-making; however, these systems also replicate and intensify existing social hierarchies through historically embedded biases in training data and classification architectures [1, 2, 3]. This practice-based doctoral research introduces *socio-technical data literacy* as a critical and reflexive framework for engaging with these systems, one that equips non-technical users to decode, interrogate, and intervene in the classificatory logics that shape generative visual outputs. Socio-technical data literacy is defined here as a situated capacity to critically engage with the interconnected technical and socio-cultural dimensions of generative AI.

The term *socio-technical* underscores that these systems are not purely computational but are shaped by cultural assumptions, institutional structures, and power relations. *Data* refers not only to training corpora but also to the classification schemas that organize and give meaning to information. These classifications—often hidden behind polished outputs—play a crucial role in determining what is made visible, normative, or excluded. *Literacy*, in this context, is not merely a technical skill but a relational, critical ability to interpret, question, and reshape algorithmic representations in a situated context. It enables users to surface bias, negotiate meaning, and collectively reimagine the epistemic structures embedded in generative systems.

Rather than treating users as passive recipients of generative technologies, this work emphasizes their role as epistemic agents with ethical responsibility in shaping model behavior through interactions such as prompt design, content selection, and sense-making. Within this perspective, the research investigates how non-technical stakeholders can actively challenge dominant representations and co-construct alternative classificatory logics within generative AI systems.

The work is guided by two central questions: How can socio-technical data literacy function as a creative-critical method for decoding and intervening in bias with generative AI systems, while supporting user agency in the interpretation and manipulation of classification processes? How can

CHItaly 2025: Technologies and Methodologies of Human-Computer Interaction in the Third Millenium, Doctoral Consortium, 6-10 October 2025, Salerno, Italy

antonella.autuori@supsi.ch (A. Autuori)
0000-0002-5725-8446 (A. Autuori)

participatory approaches to data and AI move beyond technical performativity to foster more relational, care-oriented engagements with generative technologies and data?

Informed by feminist epistemologies [4, 5], critical pedagogy [6, 7], and critical data studies [8, 9], this research approaches knowledge as situated, relational, and materially embedded. Feminist theory rejects the notion of disembodied objectivity, emphasizing instead partial perspectives grounded in lived experience and conditioned by specific socio-cultural contexts. This lens supports an understanding of human–AI interaction as mediated by both personal and structural factors, such as identity, affect, memory, language, and access. Critical pedagogy reinforces this perspective by framing learning as a dialogic and collective process, oriented toward reflection, agency, and the disruption of dominant epistemic hierarchies. Critical data studies further extend this approach by interrogating how data infrastructures embed social assumptions and reproduce asymmetries of representation. In addition to this theoretical foundation, speculative design [10] can offer methodological strategies that enable participants to challenge normative data logics and imagine alternative engagements with generative technologies.

Situated within HCI, this research advances a set of tools designed to evaluate and inform non-technical users' engagement with generative AI systems, with a focus on fostering agency and critical awareness. As these systems become increasingly embedded in everyday contexts, understanding how users interpret, challenge, and influence generative outputs becomes essential.

Designed for educational and civic settings, these tools support reflective engagement and promote individual and collective responsibility. By foregrounding user agency in shaping model behavior, the framework invites critical attention to accountability, transparency, and the classificatory systems that underpin algorithmic outputs. It thus expands the scope of data literacy toward more equitable, situated, and ethically responsive human–AI relations.

This research is currently at the beginning of its second year, with the methodology defined and under implementation. Ethics approval has been obtained from the RMIT University Human Research Ethics Committee in Melbourne, and expert interviews and participatory workshops are currently underway. The following sections outline the theoretical foundations, participatory methodology, and practical contributions of this research.

2. Research Background

The ontological foundation of this practice-based PhD engages with interdisciplinary discourses from philosophy, history, and science and technology studies to examine how technological systems, bias, and discrimination are co-produced. Understanding AI as a material and discursive infrastructure—shaped by practices of classification, representation, and decision-making—requires a historically and conceptually grounded perspective [2, 11].

Recent researches demonstrate that both large language models (LLMs) and text-to-image (TTI) systems systematically reproduce and amplify stereotypes related to gender, sexuality, and ethnicity [12, 13, 14], marginalize non-Western epistemologies [15], and constrain identity representations [16], while ultimately excluding disability and neurodivergent identities [17]. These patterns of exclusion are deeply embedded in the construction of training datasets, where selection, annotation, and classification practices are shaped by assumptions about what and who should be made visible. Such processes confer epistemic legitimacy on particular worldviews, embedding them within algorithmic systems under the guise of technical objectivity.

Data annotation has been shown to be a situated and power-laden process that mediates subjectivity and institutional authority [18, 19]. Visual taxonomies, such as those found in ImageNet, have been shown to rely on normative assumptions about what identities, roles, and expressions should look like, reducing individuals to predefined labels that claim universality but reflect narrow, culturally specific worldviews [3]. This reductive logic presumes a direct, stable correspondence between appearance and meaning, reinforcing cultural stereotypes under the guise of computational legibility. ImageNet, in particular, exemplifies the risks of large-scale annotation when applied to human subjects, where labels such as “loser,” “kleptomaniac,” or “slattern” were attached to images of real people without consent or contextual nuance [3].

As argued by Bowker and Star, classification systems embody institutional and cultural logics that, although often obscured, carry significant epistemic and political consequences [20]. In this sense, datasets do not merely reflect reality, but actively construct it through the worldviews embedded in their structures and labeling practices [21]. The act of selecting, labeling, and categorizing images is not neutral or technical—it is a political intervention with lasting impact on how people are seen, sorted, and acted upon by AI systems. These classificatory regimes not only reproduce harm but have become increasingly opaque as commercial AI systems scale, limiting public scrutiny of how representations are produced and deployed.

This research contributes to ongoing debates by framing data literacy as a critical, collective practice specific to the context of generative AI. It addresses the epistemic and political dimensions of classification systems, proposing tools and methods that make these structures accessible and open to contestation by non-technical audiences.

3. Towards a Bottom-Up Participatory Imperative

A central ambition of this research is to involve non-technical experts, community members, educators, activists, and other stakeholders who have been historically marginalized or excluded from the design and classification processes underpinning AI systems.

These individuals are not merely end-users; they are co-constructors of knowledge [22, 23] whose lived experiences, values, and perspectives are essential for surfacing biases, contesting dominant narratives, and envisioning alternative futures for AI [24].

Conventional participatory practices in AI often confine stakeholder involvement to consultative or tokenistic roles, where input is solicited only at discrete stages—typically after key design decisions have already been made—or restricted to superficial aspects such as user interface feedback [25]. Moreover, current practice frequently relies on proxies—such as UX professionals or algorithmic models—to represent stakeholder voices, rather than enabling direct and sustained engagement [26].

This results in constrained agency and limited influence over foundational classificatory structures. Such forms of engagement are not merely desirable but constitute necessary conditions for ensuring accountability, transparency, and contextual relevance in AI systems. Within this framework, the redistribution of epistemic authority and the collective shaping of classification processes are understood as central to advancing more just, reflexive, and socially responsive technological futures [27, 28, 26, 29].

This requires politically informed understandings of how technology and citizenship are entangled, making visible the power relations embedded in digital systems and supporting emancipatory practices aimed at social justice[30]. As D’Ignazio and Klein emphasize in their framework of data feminism, Expanding who participates in the design and interpretation of data is not simply a matter of broadening access, but a deliberate epistemic choice—one that challenges dominant knowledge systems and affirms the value of situated, plural forms of understanding that are often excluded from mainstream data practices [31].

4. Methodology

This PhD research adopts a practice-based methodology that weaves together critical theoretical inquiry and participatory experimentation. The approach is structured around four interconnected methodological pillars, each designed to foreground the epistemic and ethical complexities of generative AI while centering the agency of non-technical stakeholders.

4.1. Literature Review as Critical Infrastructure Mapping

The first pillar consists of a literature review conceived as a form of critical infrastructure mapping. Rather than summarizing existing work, it delineates the conceptual landscape of data classification, representational bias, and user agency within generative AI. These systems are approached as socio-technical infrastructures shaped by historical, cultural, and political conditions.

As part of this phase, a cross-disciplinary *bias cartography* is assembled—drawing from media studies, data science, informatics, and psychology—not to isolate technological failures, but to examine how algorithmic and human biases intersect. This mapping supports a relational understanding of classification processes and provides a shared foundation for reflection and discussion with experts and participants in following methods.

4.2. Expert Interviews

The second pillar involves semi-structured interviews with designers, educators, AI practitioners, and activists engaged in critical work on algorithmic systems. Conceived as dialogical rather than extractive, these interviews support knowledge co-production through reflective prompts and in-depth discussion.

Participants are invited to critically engage with and contribute to the cross-disciplinary bias cartography developed during the literature review, bringing insights from their respective domains of practice. This process surfaces tensions between ethical commitments and technical constraints, and informs the iterative co-design of workshop formats and the development of evaluation criteria.

4.3. Participatory Workshops

The third methodological pillar centers on participatory workshops, which are conceived as epistemic interventions and critical making spaces. These workshops invite participants to engage with generative AI through a combination of reflective inquiry and speculative experimentation.

Activities include prompt hacking—where participants iteratively test and annotate generative models to reveal hidden biases and zine-making, which draws on feminist traditions of storytelling to document personal encounters with algorithmic classification. Further, dataset remixing enables the co-creation of alternative taxonomies through collaborative data curation, while image annotation and reclassification exercises encourage participants to challenge normative visual grammars and disrupt established hierarchies.

These workshops are guided by principles of constructionism [32], care ethics [33], and pedagogical co-production [6], emphasizing hands-on, embodied critique over abstract deliberation. Data generated from observations, participant-created artifacts, and reflective discussions are analyzed using reflexive thematic analysis.

4.4. Toolkit and Evaluative Framework for Socio-Technical Data Literacy in Participatory and Educational Environment

The final phase of the research involves the development and assessment of a modular toolkit designed to support the practice of socio-technical data literacy in educational and civic contexts. The toolkit includes adaptable facilitation formats, design probes, and guiding materials, and is examined through an evaluative lens that attends to multiple dimensions of participant engagement.

These include epistemic understanding—the ability to articulate and identify bias in generative AI systems; social engagement—reflected in the willingness to discuss and intervene in classification processes; and critical awareness—the recognition of AI systems as situated, value-laden infrastructures.

Evaluation is not treated as a conclusive step, but rather as an iterative and generative moment within the research, one that reflects on how theoretical commitments are translated into situated practice, and how collective inquiry can inform more inclusive and contextually grounded ways of engaging with AI.

5. Conclusion and Research Contribution

This practice-based doctoral research contributes to HCI by reframing AI literacy through a socio-technical lens that centers interpretive agency, ethical responsibility, and participatory engagement. Generative AI is approached not as a neutral infrastructure but as a site where bias is embedded, enacted, and contested through interaction.

The development of a socio-technical data literacy framework enables non-technical stakeholders—often excluded from processes of design, governance, and interpretation—to critically engage with the classificatory logics shaping generative outputs. The project investigates how engagement with generative AI can support a shift from bias awareness to the cultivation of ethical agency. It focuses on how individuals recognize and respond to algorithmic representations, and how moments of interpretation can become sites of negotiation, reflexivity, and shared accountability. Bias is understood not only as a property of data but as something reproduced through interaction, sense-making, and institutional context.

Through participatory workshops, dialogical interviews, and critical making practices, the research explores how collective, situated interventions can foster more inclusive and pluralistic approaches to knowledge production. The resulting toolkit and evaluative framework offer practical resources for educational and civic contexts, while proposing new ways of assessing agency, awareness, and engagement in AI-mediated environments. Ultimately, this work expands the scope of data literacy and participatory design toward more equitable, reflexive, and socially responsive human–AI interactions.

Acknowledgments

The author would like to thank their supervisors, Brad Haylock and Laurene Vaughan, from the RMIT Design School for their unwavering support and for guiding the development of this research.

Declaration on Generative AI

During the preparation of this work, the author used Grammarly in order to: Grammar and spelling check. Further, the author used Deeply in order to: Syntax review.

References

[1] M. Broussard, More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech, MIT Press, 2023.

[2] K. Crawford, Atlas of AI: Power, politics, and the planetary costs of artificial intelligence, Yale University Press, 2021.

[3] K. Crawford, T. Paglen, Excavating ai: The politics of images in machine learning training sets, https://excavating.ai/, 2020.

[4] D. Haraway, Situated knowledges: The science question in feminism and the privilege of partial perspective, Feminist Studies 14 (1988). doi:10.2307/3178066.

[5] S. Harding, Subjectivity, experience and knowledge: An epistemology from/for rainbow coalition politics, Development and Change 23 (1992).

[6] P. Freire, Pedagogy of the oppressed, Seabury Press, 1970.

[7] B. Hooks, Teaching to transgress: Education as the practice of freedom, Routledge, 1994.

[8] J. Gray, C. Gerlitz, L. Bounegru, Data infrastructure literacy, Big Data & Society 5 (2018). doi:10.1177/2053951718786316.

[9] L. Pangrazio, N. Selwyn, Critical Data Literacies: Rethinking Data and Everyday Life, MIT Press, 2023.

[10] N. Sánchez Querubín, S. Niederer, Climate futures: machine learning from cli-fi, Convergence 30 (2024). doi:10.1177/13548565221135715.

[11] G. Simondon, Du mode d'existence des objets techniques, Aubier, 1958.

[12] A. S. Luccioni, C. Akiki, M. Mitchell, Y. Jernite, Stable bias: Analyzing societal representations in diffusion models, arXiv:2303.11408, https://arxiv.org/abs/2303.11408, 2023.

[13] L. Nicoletti, D. Bass, Humans are biased. generative ai is even worse, Bloomberg Technology + Equality, https://www.bloomberg.com/graphics/2023-generative-ai-bias, 2023.

[14] UNESCO, Generative ai: Unesco study reveals alarming evidence of regressive gender stereotypes, https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes, 2024.

[15] V. Turk, How ai reduces the world to stereotypes, Rest of World, https://restofworld.org/2023/ai-image-stereotypes/, 2023.

[16] A. F. d. C. Vázquez, E. C. Garrido-Merchán, A taxonomy of the biases of the images created by generative artificial intelligence, arXiv:2407.01556, https://doi.org/10.48550/arXiv.2407.01556, 2024.

[17] Y. Welker, Algorithmic diversity: Mitigating ai bias and disability exclusion, Forbes, https://www.forbes.com/councils/forbestechcouncil/2023/05/09/algorithmic-diversity-mitigating-ai-bias-and-disability-exclusion/, 2023.

[18] M. Miceli, M. Schuessler, T. Yang, Between subjectivity and imposition: Power dynamics in data annotation for computer vision, Proc. ACM Hum.-Comput. Interact. 4 (2020). doi:10.1145/3415186.

[19] A. Arzberger, S. Buijsman, M. L. Lupetti, A. Bozzon, J. Yang, Nothing comes without its world – practical challenges of aligning llms to situated human values through rlhf, in: Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society, AIES '24, AAAI Press, San Jose, CA, USA, 2024. doi:10.1609/aies.v7i1.31617.

[20] G. C. Bowker, S. L. Star, Sorting things out: Classification and its consequences, MIT Press, 1999.

[21] H. Davis, A dataset is a worldview, Medium, https://medium.com/data-science/a-dataset-is-a-worldview-5328216dd44d, 2019.

[22] C. DiSalvo, Design and the construction of publics, Design Issues 25 (2009). doi:10.1162/desi.2009.25.1.48.

[23] A. K. Munk, V. Tommaso, A. Meunier, Data sprints: A collaborative format in digital controversy mapping, in: J. Vertesi, D. Ribes (Eds.), Digital STS: A field guide for science & technology studies, Princeton University Press, 2019. doi:10.2307/j.ctvc77mp9.34.

[24] S. Costanza-Chock, Design justice: Community-led practices to build the worlds we need, MIT Press, 2020.

[25] F. Delgado, S. Yang, M. Madaio, Q. Yang, The participatory turn in ai design: Theoretical foundations and the current state of practice, in: EAAMO '23: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, ACM, 2023. doi:https://doi.org/10.1145/3617694.3623261.

[26] M. Sloane, E. Moss, O. Awomolo, L. Forlano, Participation is not a design fix for machine learning, in: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization EAAMO '22:, ACM, 2020. doi:https://doi.org/10.1145/3551624.3555285.

[27] HEK House of Electronic Arts Basel, Participatory ai: How to make better ai, https://share.hek.ch/en/participatory-ai-how-to-make-better-ai, 2024.

[28] D. J. Nucera, V. Mogilevich, Teaching Community Technology Handbook, Detroit Community Technology Project, 2015.

[29] P. Gourlet, D. Ricci, M. Crépel, Reclaiming artificial intelligence accounts: A plea for a participatory turn in artificial intelligence inquiries, Big Data & Society April-June (2024). doi:10.1177/20539517241248093.

[30] A. Emejulu, C. McGregor, Towards a radical digital citizenship in digital education, Critical Studies in Education 60 (2019). doi:10.1080/17508487.2016.1234494.

[31] C. D'Ignazio, L. F. Klein, Data feminism, MIT Press, 2020.

[32] S. Papert, Mindstorms—Children, Computers and Powerful Ideas, Basic Books, 1980.

[33] M. Puig de la Bellacasa, Matters of care in technoscience: Assembling neglected things, Social Studies of Science 41 (2017). doi:10.1177/0306312710380301.