Beyond Templates: Understanding and Addressing Human-AI Interaction Harms through Practitioners' Assumptions

Julia De Miguel Velazquez[1]

[1]King's College London
julia.de_miguel_velazquez@kcl.ac.uk

Abstract

Human-AI interaction risks account for most real-world AI harms but remain underrepresented in safety evaluations. Instead, these tend to prioritize model-level evaluations, abstracting away the contexts in which harms emerge. In tackling this, responsible AI efforts have provided practitioners with tools, such as checklists and impact assessments. Yet, these tools often assume a shared understanding of harm, overlooking practitioners' personal, organizational, and media assumptions. As research increasingly addresses human-AI interaction risks, it is crucial to examine practitioners' assumptions. I first conduct a survey and interviews to empirically explore how practitioners envision harm through their underlying assumptions. Second, I reflect on these findings to explore how responsible AI efforts can better support critical reflection on underlying assumptions.

Introduction

> In controversies about technology and society, there is no idea more provocative than the notion that technical things have political qualities.
> *—Langdon Winner, 1980*

When teaching AI Ethics at a prominent AI organization, I noticed practitioners' concerns diverged: some emphasized artificial general intelligence and competition with China as their main concern; others argued back by pointing at more immediate risks, such as overreliance or biases. These claims often mirrored media's opposing narratives (Rothman 2025), risking the framing of risks through moral panics rather than evidence-based assessment. In this proposal I outline a plan to systematically study how practitioners' personal, organizational, and media influences shape AI safety evaluations, to identify current blind spots.

In my previous work I found most real-world AI harms (69%) occur in human-AI interactions (De Miguel Velázquez et al. 2024). However, these harms remain underrepresented in safety evaluations—making less than 10% in 2024 (Weidinger et al. 2023). Most evaluations focus on model-level failures, overlooking the complex, socially situated harms in human-AI interactions (Weidinger et al. 2023; Feng et al. 2025; Ibrahim et al. 2024). For instance, even if LLMs have an almost perfect medical score in a benchmark, this performance does not necessarily translate to user interactions, where accuracy drops (Bean et al. 2025). This gap in evaluations is a problem (see first row in Table 1).

To partially address this, responsible AI communities such as CHI, CSCW, FAccT, and AIES have made valuable progress by designing tools, such as templates or impact assessments, that help practitioners anticipate negative impacts (Deng, Barocas, and Wortman Vaughan 2025). However, these tools might fall short in their goal as they assume practitioners have a "view from nowhere"– the illusion of objective, universal knowledge that neglects the situated perspective of those producing it (Haraway 1988; Widder and Nafus 2023). In this paper, I aim to study practitioners' reactions to these impact assessments themselves and conceptualizations of harm. As social scientists highlight, AI harm conceptualizations are instead shaped by personal experiences, organizational norms, and media narratives (Elish 2019; Olson et al. 2025) – in short, shaped by practitioners *assumptions*. Therefore, I propose the following research question (RQ):

> ***RQ***: How do practitioners perceive and interpret human-AI interaction harms, given personal experiences, organizational norms, and media narratives?

This research will focus on generative AI, particularly LLMs, due to data availability, though results may extend to other modalities. To answer the *RQ*, I will conduct a user study on practitioners. Building on these insights I will reflect on the best practices of how we can design an educational guide for the research agenda with conceptual tools that enables critical and ethical thinking when designing safety evaluations. This will contribute to the growing research agenda for human-AI interaction and sociotechnical safety evaluations (Feng et al. 2025).

This proposal partially complements my broader PhD thesis on how human-AI interaction risks are under-recognized in technical discourse, which includes my previous work on real-world AI incidents (De Miguel Velázquez et al. 2024).

Related Work

This section outlines the relevant fields for this work, also summarized in Table 1.

Field	How it understands harm	Assumptions	Methods	Gap
AI and Tech Ethics	As a technical failure, bias, or un/fairness	Harms are measurable, often at the model-level	Benchmarks, red teaming, alignment evaluations	Safety efforts still neglect interactional harms arising from user-level deployment (Ibrahim et al. 2024).
Human-Computer Interaction (HCI)	As situated, relational, and experienced	Design shapes experience and agency	User studies, participatory design	Tools exist, but can benefit from assessing their efficiency (Berman, Goyal, and Madaio 2024)
Science and Technology Studies (STS)	As socially constructed and politicized	Harms emerge through power and discourse	Discourse analysis, ethnography	STS insights rarely inform concrete harm mitigation practices (Ananny and Crawford 2018)

Table 1: Background and research gap across fields. The aim of the paper is to combine human-computer interaction (second row) and STS (third row) to solve the gap from tech ethics (first row).

Do harm anticipation tools work in practice? Tools such as impact assessments can be useful for AI practitioners to uncover potential negative impacts (Deng, Barocas, and Wortman Vaughan 2025). Researchers have co-designed tools that are practical, and empirically grounded (Rakova et al. 2021a; Metcalf et al. 2021; Do et al. 2023; Deng, Barocas, and Wortman Vaughan 2025). Reflecting on how these tools work in practice, a systematic review found they are often evaluated for usability, rather than for whether they achieve their intended purpose. (Berman, Goyal, and Madaio 2024). The way the impacts are operationalized in impact assessments may diverge from actual harms, so that usability alone is insufficient. (Metcalf et al. 2021). Less attention has been given to on how practitioners respond to harms given their pre-existing social beliefs.

Organizational and media influences on harm perception Scholars like Forsythe (1993), Wajcman (2007), and Suchman (2002) show how decisions in research embed social values: biases influence workflows and design decisions, and even fairness claims within technical systems can obscure systemic injustices in practice (Selbst et al. 2019). Practitioners' understandings of harm are often shaped more by personal experience, organizational norms, and media narratives than by social science frameworks (Rakova et al. 2021b; Selbst et al. 2019). For example, practitioners' demographic backgrounds influence how they perceive and act on ethical concerns in AI (Olson et al. 2025), while accountability is also shaped by press narratives (Elish 2019).

Research Design

Methods To address the *RQ* (practitioners' interpretations of harm) I will design and conduct a set of interviews and a survey.

1. *Survey and interview design:* the survey and interview questions will be informed by a literature review on AI ethics, organizational decision-making, and media influences, as well as insights from real-world AI incidents (De Miguel Velázquez et al. 2024).
2. *Conducting survey and interviews:* the survey will be distributed online to reach a broad range of AI practitioners across sectors. The interviews will be conducted online or in person, will last around 45-60 minutes, in order to allow in-depth exploration of participants' experiences with AI harms, organizational norms, and media narratives.
3. *(Tentative) Focus groups:* focus groups could be conducted as an additional method to observe the collective discussions about harm perception. These remain tentative and will be refined based on consortium feedback.

Participants In this research, I define practitioners broadly as those who work in any role on a team that develops products or services involving AI, following other work (Holstein et al. 2019; Orr and Davis 2020). I will prioritize those whose work focuses on designing and conducting safety evaluations or responsible AI work. Sampling will be purposive, targeting influential professionals (Collett 2024), using networking, and snowballing sampling.

Analysis I will use critical discourse analysis to analyze the survey and interviews to reveal how practitioners interpret human-AI interaction harms, and to partially reveal how these interpretations shape safety evaluations (Carvalho 2008). I will extract actionable takeaways, such as patterns and gaps in understanding human-AI risks, and ways to improve impact assessments for ethical reflection.

Limitations A practical limitation will be to access practitioners which may limit sample size (Collett 2024). I would like to address this at AIES and hear from professionals on the field how to mitigate this.

Conclusion

By systematically exploring how practitioners perceive and interpret human-AI interaction harms through personal, organizational, and media norms, this research will contribute to understanding their assumptions when using impact assessment tools. The findings from the survey and interviews will uncover the assumptions practitioners bring to AI safety evaluations. These insights will reveal both blind spots and opportunities for responsible AI education. Ultimately, this work emphasizes that harm is not merely a technical failure, but a socially and politically situated phenomenon. Harm is not neutral—it is politically defined (Ananny 2024; Hilgartner and Bosk 1988).

References

Ananny, M. 2024. Making generative artificial intelligence a public problem. Seeing publics and sociotechnical problem-making in three scenes of AI failure. *Javnost-The Public*, 31(1): 89–105.

Ananny, M.; and Crawford, K. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20(3): 973–989.

Bean, A. M.; Payne, R.; Parsons, G.; Kirk, H. R.; Ciro, J.; Mosquera, R.; Monsalve, S. H.; Ekanayaka, A. S.; Tarassenko, L.; Rocher, L.; et al. 2025. Clinical knowledge in LLMs does not translate to human interactions. *arXiv preprint arXiv:2504.18919*.

Berman, G.; Goyal, N.; and Madaio, M. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–24.

Carvalho, A. 2008. Media (ted) discourse and society: Rethinking the framework of critical discourse analysis. *Journalism studies*, 9(2): 161–177.

Collett, C. 2024. The hustle: How struggling to access elites for qualitative interviews alters research and the researcher. *Qualitative Inquiry*, 30(7): 555–567.

De Miguel Velázquez, J.; Šćepanović, S.; Gvirtz, A.; and Quercia, D. 2024. Decoding Real-World AI Incidents. *IEEE Computer*, 57(11).

Deng, W. H.; Barocas, S.; and Wortman Vaughan, J. 2025. Supporting Industry Computing Researchers in Assessing, Articulating, and Addressing the Potential Negative Societal Impact of Their Work. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–37.

Do, K.; Pang, R. Y.; Jiang, J.; and Reinecke, K. 2023. "That's important, but...": How Computer Science Researchers Anticipate Unintended Consequences of Their Research Innovations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Elish, M. C. 2019. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)*.

Feng, K. J. K.; Pang, R. Y.; Kuo, T.-S.; Winecoff, A.; Tseng, E.; Widder, D. G.; Suresh, H.; Reinecke, K.; and Zhang, A. X. 2025. Sociotechnical AI Governance: Challenges and Opportunities for HCI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.

Forsythe, D. E. 1993. Engineering knowledge: The construction of knowledge in artificial intelligence. *Social studies of science*, 23(3): 445–477.

Haraway, D. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3): 575–599.

Hilgartner, S.; and Bosk, C. L. 1988. The rise and fall of social problems: A public arenas model. *American journal of Sociology*, 94(1): 53–78.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–16.

Ibrahim, L.; Huang, S.; Bhatt, U.; Ahmad, L.; and Anderljung, M. 2024. Towards interactive evaluations for interaction harms in human-AI systems. *arXiv preprint arXiv:2405.10632*.

Metcalf, J.; Moss, E.; Watkins, E. A.; Singh, R.; and Elish, M. C. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 735–746. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Olson, L.; Anna-Lena Fischer, R.; Kunneman, F.; and Guzmán, E. 2025. Who Speaks for Ethics? How Demographics Shape Ethical Advocacy in Software Development. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2847–2862.

Orr, W.; and Davis, J. L. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, 23(5): 719–735.

Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021a. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021b. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23.

Rothman, J. 2025. Two Paths for A.I. *The New Yorker*.

Selbst, A. D.; boyd, d.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

Suchman, L. 2002. Located accountabilities in technology production. *Scandinavian journal of information systems*, 14(2): 7.

Wajcman, J. 2007. From women and technology to gendered technoscience. *Information, Community and Society*, 10(3): 287–298.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Widder, D. G.; and Nafus, D. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1): 20539517231177620.