Generative AI Going Awry: Enabling Designers to Proactively Avoid It in CSCW Applications

Jed R. Brubaker
University of Colorado Boulder
Boulder, CO USA
Jed.Brubaker@Colorado.EDU

Casey Fiesler
University of Colorado Boulder
Boulder, CO USA
Casey.Fiesler@Colorado.EDU

Michael Madaio
Google Research
New York, NY USA
madaiom@google.com

John Tang†
Microsoft Research
Mountain View, CA USA
johntang@microsoft.com

Richmond Y. Wong
Georgia Institute of Technology
Atlanta, GA USA
rwong34@gatech.edu

Abstract

The rapid development and deployment of generative AI technologies creates a design challenge of how to proactively understand the implications of productizing and deploying these new technologies, especially with regard to negative design implications. This is especially concerning in CSCW applications, where AI agents can introduce misunderstandings or even misdirections with the people interacting with the agent. In this panel, researchers from academia and industry will reflect on their experiences with ideas, methods, and processes to enable designers to proactively shape the responsible design of genAI in collaborative applications. The panelists represent a range of different approaches, including speculative fiction, design activities, design toolkits, and process guides. We hope that the panel encourages a discussion in the CSCW community around techniques we can put into practice today to enable the responsible design of genAI.

CCS Concepts

• **Human-centered computing** → **Collaborative and social computing**; *Collaborative and social computing design and evaluation methods*

Keywords

Generative AI, design, redteaming

†Contact author

CSCW Companion '24, November 9-13, 2024, San Jose, Costa Rica.

ACM ISBN 979-8-4007-1114-5/24/11.
https://doi.org/10.1145/3678884.3689133

ACM Reference format:

Jed R. Brubaker, Casey Fiesler, Michael Madaio, John Tang, and Richmond Y. Wong. 2024. Generative AI Going Awry: Enabling Designers to Proactively Avoid It in CSCW Applications. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24), November 9-13, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 4 pages, https://doi.org/10.1145/3678884.3689133

1 Introduction

With the meteoric rise of generative AI, researchers and developers in academia and industry have been rushing to explore how genAI can be applied to our work. A recent New York Times article described how Satya Nadella, CEO of Microsoft, "...told all of his lieutenants to find ways to build A.I. into Microsoft's many, many products, even though the technology didn't always work correctly" [13]. In this competitive rush to include genAI features into our technology, it is hard to account for or anticipate negative design implications of incorporating genAI into our work practices. Marchal, et al. [8] documented the ways genAI has been misused with malicious intent after deployment. However, we need to help designers anticipate negative design implications of genAI early in the design process, so that they can proactively develop mitigations and guardrails **before** the technology gets deployed. And besides maliciously abusing genAI, there are also unintended negative consequences of using genAI that may be less obvious but still important to avoid.

We think it is timely to share ideas and experiences of how to enable design teams to anticipate potential "dark" implications of AI technology to encourage a responsible design process in developing the technology before it gets deployed. The panelists represent people from both academia and industry who will share their perspectives on ideas, methods, or experiences to proactively design AI responsibly. We wanted to share these perspectives with the CSCW community to help encourage and accelerate these responsible design efforts to try to keep pace with the rush to develop and deploy genAI.

The panel consists of colleagues who are actively involved in efforts to shape the responsible design process of genAI. Each panelist will describe their experiences as a way of encouraging

discussion among our community on how to encourage more responsible design of genAI.

2 Panelists

2.1 Casey Fiesler

Casey Fiesler is an Associate Professor in Information Science at University of Colorado Boulder. One component of her work is the creation of tools and educational interventions towards reducing "ethical debt" in technology design and research [1] by encouraging forethought about unintended consequences. For example, her "Black Mirror Writers Room" exercise encourages creative speculation about what might go wrong with technology in the future, towards developing ethical speculation as a skill that could be applied to current projects [2]. Though originally developed for classroom use, the exercise has also been used in research and industry settings. Fiesler also conducts research into computing ethics education, particularly towards strategies for encouraging students to think about ethical implications throughout every stage of technology design and implementation [10].

Creator Charlie Brooker once said of *Black Mirror* that the show doesn't tend to be about the technology itself, but rather how we use or misuse it. This context, and the framing of the exercise itself, encourages speculation about complex sociotechnical harms: not what might go wrong with the technology itself, but the consequences of its embedding into society. In recent years, groups of both students and practitioners engaging in this exercise have frequently focused on AI and in particular, core problems of interest to the CSCW community: e.g., facial recognition bias [9], AI clones or misinformation [3], or AI in healthcare [18].

2.2 John Tang

John Tang is a Senior Principal Researcher at Microsoft Research. His recent research has focused on developing genAI-powered agents to support collaboration. We have been exploring the concept of Dittos–mimetic agents that look and sound like you that can represent you in meetings that you cannot attend [4]. We realized that this is a provocative concept, so we wanted to explore ways in which Dittos could go "dark" to shape our design process while it is still in the formative stage. By now, we have held two "Dark Ditto" workshops where we gathered the research and engineering team (including interns) to focus on negative design implications of Dittos. Working through exercises inspired by the Black Mirror Writer's Room work [2], we brainstormed concerns and developed scenarios illustrating ways in which Dittos in social use among teams of people could stray from their design intent. These workshops were not only meaningful learning experiences for our teams, but also raised awareness and helped shape the design and development of Dittos while the concept was still in formation.

For example, one scenario speculated about what would happen if an employee violated the policy of limiting only one Ditto per person. By sending out an army of Dittos to attend a wide range of meetings, he could consolidate information in ways that was not humanly possible. That accumulation of information led to great organizational power, plus the realization that he did not need as many employees to get the job done. This scenario not only demonstrated how the usage of Dittos could become problematic, but also the importance of developing **policies of usage**, which go beyond responsible design. While a company could invest great efforts in designing technology responsibly, the development of policies around how that technology is used is largely in the hands of the customers who purchase and deploy that technology. This scenario illustrates a major concern in the deployment of genAI technologies into use.

2.3 Richmond Y. Wong

Richmond Wong is an Assistant Professor of Digital Media at the Georgia Tech School of Literature, Media, and Communication. His research involves creating design tools and processes that can help technologists identify and discuss social values and ethical issues in organizational work contexts. These activities draw on practices of speculative design and design fiction to try to foresee potential social harms before they occur, often focusing on contested social values and the role of sociotechnical infrastructures.

One design activity, design fiction workbooks, uses the creation of fictional product scenarios to help surface discussion of multiple conceptions of the same social value, showing how privacy is conceptualized differently depending on the **socio-technical relationships and contexts** at play [15]. For example, one product was imagined to be deployed as an information sharing service in two contexts: between intimate partners, and in the workplace to mediate employer-employee relationships. With the design fictions, participants were able to discuss how privacy was conceptualized differently among each of these types of interactions and relationships. The activity can be adapted to similarly discuss how social values identified as important to responsible AI—such as fairness or transparency—are conceptualized differently when a system is used within different types of social relationships.

A second design activity, *Timelines*, focuses on exploring potential long-term secondary and tertiary social and ethical effects of technologies through the creation of fictional headlines and social media posts [16]. This activity draws attention to the role of **socio-technical infrastructures** that can create long-term and shared impacts among diverse stakeholder communities, rather than focusing on short-term individual experiences. This activity has been used in prior CSCW workshops [11, 17] and can be adapted to think about the socio-technical infrastructures related to generative AI.

2.4 Michael Madaio

Michael Madaio is a Senior Research Scientist at Google Research. His research draws on methods from human-computer interaction to develop and study tools and processes to support proactive work to anticipate and address potential societal impacts of AI. In one line of work, he has co-designed a process

guide with AI practitioners to proactively identify and assess potential fairness issues in AI development [5], and empirically studied how AI teams assess fairness impacts [6] and customize a general-purpose fairness process for their specific applications and use cases—a challenge exacerbated by the design paradigm of pre-trained generative AI models [7].

In another line of work, he has studied how AI practitioners learn on-the-job about responsible AI, including what concepts and skills they are learning, via what pathways, identifying a need to foster AI practitioners' ability to proactively anticipate potential downstream harms [7]. To address this gap, he and his collaborators developed an *in situ* tool to support harm envisioning during the prototyping phase of LLM application design [12].

However, a recurring theme across all of the aforementioned studies of responsible AI tools and processes is how responsible AI design introduces new forms of collaborative work among practitioners, mediated by their organizational contexts. As such, the CSCW community is well-positioned to contribute to this emerging area of research into the proactive design of AI.

3 Panel Structure

The panel will be moderated by Jed R. Brubaker. Each panelist will have 7 minutes to present a position statement describing their experiences with methods, design exercises, and other ideas for enabling designers to anticipate how genAI could go awry and mitigating that in the design of collaborative technologies. The goal is to give our community concrete ideas to incorporate in their own design processes for developing collaborative genAI technology. After going through each panelist's position statement, we will discuss some questions shared in advance, such as what their biggest design concern with genAI is in collaborative applications, and what one thing people in the audience can do in their own design work. Then we will open the panel up to questions from the audience.

Acknowledgements

We thank the many participants who have been involved in the various exercises and activities described in this panel. Jeb Brubaker acknowledges funding from NSF #2048244.

References

[1] Casey Fiesler. AI has social consequences, but who pays the price? Tech companies' problem with 'ethical debt.' The Conversation. https://theconversation.com/ai-has-social-consequences-but-who-pays-the-price-tech-companies-problem-with-ethical-debt-203375. 2023.

[2] Shamika Klassen and Casey Fiesler (2022). "Run Wild a Little With Your Imagination": Ethical Speculation in Computing Education with Black Mirror, *SIGCSE 2022: Proceedings of the 53rd ACM Technical Symposium on Computer Science Education*, February 2022, pp. 836 - 842 https://doi.org/10.1145/3478431.3499308

[3] Patrick Yung Kang Lee, Ning F. Ma, Ig-Jae Kim, and Dongwook Yoon (2023). Speculating on risks of AI clones to selfhood and relationships: Doppelganger-phobia, identity fragmentation, and living memories. *Proceedings of the ACM on Human-Computer Interaction* 7, no. CSCW1 (2023) 1-28.

[4] Joanne Leong, John Tang, Edward Cutrell, Sasa Junuzovic, Greg Baribault, Kori Inkpen (2024). Dittos: Personalized, Embodied Agents that Participate in Meetings When You are Unavailable, *CSCW 2024* in press.

[5] Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

[6] Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., & Wallach, H. (2022). Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), 1-26.

[7] Madaio, M. A., Chen, J., Wallach, H., & Wortman Vaughan, J. (2024). Tinker, Tailor, Configure, Customize: The Articulation Work of Contextualizing an AI Fairness Checklist. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-20.

[8] Nahema Marchal, Rachel Xu, Rasmi Elasmar, Iason Gabriel, Beth Goldberg, William Isaac. Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data, https://arxiv.org/abs/2406.13843 (accessed July 12, 2024).

[9] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker (2019). How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (2019) 1-33.

[10] Michelle Tran & Casey Fiesler (2024). "It's Not Exactly Meant to Be Realistic": Student Perspectives on the Role of Ethics in Computing Group Projects. *Proceedings of the ACM ICER International Computing Education Research Conference* 2024.

[11] Jessica Vitak, Michael Zimmer, Anna Lenhart, Sunyup Park, Richmond Y. Wong, and Yaxing Yao (2021). Designing for Data Awareness: Addressing Privacy and Security Concerns About "Smart" Technologies. *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21 Companion).* 2021. 364–367. https://doi.org/10.1145/3462204.3481724

[12] Z. J. Wang, C. Kulkarni, L. Wilcox, M. Terry, and M. Madaio (2024). Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024, May). pp. 1-40.

[13] Karen Weise & Cade Metz (2024). How Microsoft's Satya Nadella Became Tech's Steely Eyed A.I. Gambler, *New York Times.* July 14, 2024 https://www.nytimes.com/2024/07/14/technology/microsoft-ai-satya-nadella.html.

[14] Richmond Y. Wong, Andrew Chong, and R. Cooper Aspegren (2023). Privacy Legislation as Business Risks: How GDPR and CCPA are Represented in Technology Companies' Investment Risk Disclosures. *Proc. ACM Human-Computer. Interaction* 7, CSCW1, Article 82 (April 2023), 26 pages. https://doi.org/10.1145/3579515

[15] Richmond Y. Wong, Deirdre K. Mulligan, Ellen Van Wyk, James Pierce, and John Chuang (2017). Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks. *Proc. ACM Human-Computer. Interaction.* 1, CSCW, Article 111 (November 2017), 26 pages. https://doi.org/10.1145/3134746

[16] Richmond Y. Wong and Tonya Nguyen (2021). Timelines: A World-Building Activity for Values Advocacy. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Article 616, 1–15. https://doi.org/10.1145/3411764.3445447

[17] Yaxing Yao, Richmond Wong, Pardis Emami-Naeini, Nick Merrill, Xinru Page, Yang Wang, and Pamela Wisniewski (2019). Ubiquitous Privacy: Research and Design for Mobile and IoT Platforms. *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '19 Companion).* 533–538. https://doi.org/10.1145/3311957.3359430

[18] Xiao Zhan, Noura Abdi, William Seymour, and Jose Such (2024). Healthcare Voice AI Assistants: Factors Influencing Trust and Intention to Use. *Proceedings of the ACM on Human-Computer Interaction* 8, no. CSCW1 (2024): 1-37.