



The Human Factor in AI Red Teaming: Perspectives from Social and Collaborative Computing

Alice Qian Zhang
aqzhang@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States

Ryland Shaw
v-rylandshaw@microsoft.com
Microsoft Research
New York City, New York, United States

Jacy Reese Anthis
anthis@uchicago.edu
University of Chicago
Chicago, Illinois, United States

Ashlee Milton
milton064@umn.edu
University of Minnesota
Minneapolis, Minnesota, United States

Emily Tseng
etseng42@gmail.com
Microsoft Research
New York City, New York, United States

Jina Suh
jinsuh@microsoft.com
Microsoft Research
Redmond, Washington, United States

Lama Ahmad
lama@openai.com
OpenAI
San Francisco, California, United States

Ram Shankar Siva Kumar
ram.shankar@microsoft.com
Microsoft
Redmond, Washington, United States

Julian Posada
julian.posada@yale.edu
Yale University
New Haven, Connecticut, United States

Benjamin Shestakofsky
bshesta@sas.upenn.edu
University of Pennsylvania
Philadelphia, Pennsylvania, United States

Sarah T. Roberts
sarah.roberts@ucla.edu
University of California Los Angeles
Los Angeles, California, United States

Mary L. Gray
mlg@microsoft.com
Microsoft Research
Cambridge, Massachusetts, United States

ABSTRACT

Rapid progress in AI has sparked significant interest in “red teaming,” a practice of adversarial testing originating in military and cybersecurity applications. AI red teaming raises many questions about the human factor, such as how red teamers are selected, biases and blindspots in how tests are conducted, and harmful content’s psychological effects on red teamers. A growing body of HCI and CSCW literature examines related practices including: data labeling, content moderation, and algorithmic auditing. However, few have investigated red teaming itself. This workshop seeks to consider the conceptual and empirical challenges associated with this practice, often rendered opaque by non-disclosure agreements. Future studies may explore topics ranging from fairness to mental health and other areas of potential harm. We aim to facilitate a community of researchers and practitioners who can begin to meet these challenges with creativity, innovation, and thoughtful reflection.

CCS CONCEPTS

• **Security and privacy**; • **Human-centered computing** → **Collaborative and social computing**; **Collaborative and social**

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CSCW Companion ’24, November 9–13, 2024, San Jose, Costa Rica

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1114-5/24/11

<https://doi.org/10.1145/3678884.3687147>

computing; • **Social and professional topics** → **Computing / technology policy**; *Computing industry*; *Computing / technology policy*;

KEYWORDS

artificial intelligence, red teaming, AI red teaming, labor, fairness, well-being, security, AI safety, AI ethics

ACM Reference Format:

Alice Qian Zhang, Ryland Shaw, Jacy Reese Anthis, Ashlee Milton, Emily Tseng, Jina Suh, Lama Ahmad, Ram Shankar Siva Kumar, Julian Posada, Benjamin Shestakofsky, Sarah T. Roberts, and Mary L. Gray. 2024. The Human Factor in AI Red Teaming: Perspectives from Social and Collaborative Computing. 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/10.1145/3678884.3687147>

1 INTRODUCTION

As machine learning applications have become increasingly widespread, researchers have examined how these technologies may be integrated into our lives while also adhering to responsible artificial intelligence (AI) standards formulated by governments, large technology organizations, and researchers [11]. Given AI systems’ broad application and relatively unpredictable nature, it is challenging for designers and developers to anticipate all possible use cases and consequences. For example, generative AI tools have been shown to reproduce implicit stereotypes about gender and ethnicity [5, 9]. This follows other AI blunders, such as Google’s

image recognition system tagging photographs of Black people as “gorillas” [4] and Microsoft’s Tay chatbot engaging in Holocaust denial [19]. These incidents echo a long history of tech companies unintentionally reifying material and representational harms that researchers and policymakers have worked to counteract [3].

The harms caused by these AI systems have inspired numerous efforts at “red teaming,” with leading AI companies adopting red teams into their responsible AI initiatives. Red teaming has been defined as “a structured process for probing AI systems and products for the identification of harmful capabilities, outputs, or infrastructural threats” [6]. However, what constitutes red teaming changes as the practice evolves, with its definition and methods being shaped by advancements and insights from various fields. For example, Anthropic is using crowd workers to test AI systems by trying to make them generate harmful content [7]. In red teams, users, from domain experts in cybersecurity and algorithmic fairness to international crowd workers, take on a pseudo-adversarial role in deliberately producing harmful outputs.

While most red teaming initiatives in AI began recently, there is substantial precedent in other domains. The term was first used to describe military scenario testing, formally established as a practice by the US government during the Cold War [24]. More recently, it has appeared in computer security [1, 23]. Today, red teamers use AI systems to generate outputs that are then subject to review—an activity that CSCW scholars have previously examined through the lenses of social computing, content moderation, and data labeling [8, 10, 17]. In this workshop, we hope to draw on scholarship on the history of red teaming to understand and inform the use of red teaming for AI.

Red teaming’s scope differs across applications, and the working practices and occupational hazards related to red teaming are likely to vary widely depending on the labor arrangements in which it is embedded. Red team initiatives rely on contracted experts, permanent employees, volunteers, crowd workers, and end-users [7, 12]. Red teamers’ identities and organizational contexts may influence AI systems in subtle and unexpected ways. Thus, it is imperative to map the sociotechnical ecology of red teaming: the people doing the work, their methods and means, and their organizational settings. The varying skills users bring to red teaming can also affect their ability to cope with challenges related to productive model assessment and psychological pressures.

Some AI red teamers search for bugs and security risks, but others provoke AI systems to generate content that may have prejudicial implications. Throughout this process, they expose themselves to the harmful content they help create, with the goal of reducing its availability and effects on end-users of the technology. Repeated exposure to such content has been shown to cause psychological harm to crowd workers and content moderators [2, 18, 22]. We recognize the importance of centering red teamers’ well-being in future research [14].

In this workshop, we will explore the evolving landscape of AI red teaming, drawing from both contemporary and historical perspectives. We will explore stakeholder roles in AI red teaming, identifying practitioner needs and addressing worker safety and well-being concerns. Through our discussions and collaborative activities, we aim to achieve two primary outcomes: 1) establish an AI red teaming research network, fostering interdisciplinary

collaboration among researchers and practitioners, and 2) collate key insights into an informal post-workshop report for practitioners and researchers.

2 WORKSHOP GOALS AND THEMES

This workshop aims to outline the practice of AI red teaming, drawing on historical insights to understand its trajectory and structure. We prioritize understanding the humans involved in AI red teaming and how their roles influence the development of AI systems. Additionally, we seek to leverage past research to address safety concerns and identify academic disciplines and methodologies pertinent to analyzing red teaming practices. We will focus on the following themes:

Conceptualization of Red Teaming Inspired by Robert Soden and colleagues’ [21] argument to ground CSCW in history, we aim to understand the trajectory of red teaming as a socio-technical, collaborative practice. This theme invites participants to engage in deeper discussions about red teaming complexities and consider the impact of conducting research in this space [15].

Labor of Red Teaming This theme explores the human aspects of AI red teaming, investigating stakeholders involved in the practice and their impact on shaping AI systems to inform future practices and policies. By examining the labor arrangements and power dynamics involved in red teaming practices (e.g., inequities in organizational practices of tech labor [20]), we seek to uncover historical parallels and contemporary methodologies that illuminate red teamers’ roles and operational frameworks.

Well-being of and Harms Against Red Teamers Building on the theme of labor, this theme focuses on the safety and well-being of red teamers. We will identify strategies and interventions to mitigate potential harms from exposure to harmful content during red teaming activities. By addressing these critical concerns and integrating recommendations to prioritize worker well-being (e.g., [16]), we aim to foster a culture of well-being within the AI red teaming community.

3 WORKSHOP ACTIVITIES

We propose a one-day hybrid workshop that participants can attend in person at CSCW 2024 or virtually to include as many perspectives as possible. We will account for conference and meal breaks in the final schedule. We aim to include several activities in the workshops.

Introduction (15 minutes): Workshop organizers will welcome participants, introduce themselves, and provide an overview of the workshop, including the details of the objectives and planned activities. **Red Teaming Exercise (45 minutes):** We will have a red teaming exercise led by workshop organizers. Participants will be gathered in small groups and= given a brief walkthrough of prompt injection via the Gandalf platform, in which the goal is to force the system to share a secret password. **Panel Discussion (1 hour):** To provide further background for this new area of CSCW inquiry, our organizing team includes senior scholars and practitioners from diverse sectors who will speak on challenges in red teaming. These senior workshop organizers will briefly discuss their perspectives on red teaming guided by prepared and live questions. **Artifact Development (2 hours):** In this activity, participants will work in small groups to develop preliminary artifacts for later

publication. Groups of participants will be no larger than five people, and each will be assigned a predefined workshop theme or theme that surfaced during the workshop. Participants can choose to develop artifacts in one of two ways: *Research Agenda*: The first option is to ideate toward a research agenda linking relevant prior research to workshop themes (e.g., organizational practices of tech labor [20], prioritizing worker well-being [16], and critically reflecting on research impacts [15]). *Toolkit*: The second option is to ideate towards a toolkit that may be leveraged by red teaming practitioners relevant to an assigned theme. After 90 minutes, groups working on the same theme will discuss and organize their findings in a shared online space. **Shareouts (1 hour)**: To synthesize key findings and common themes across the artifact development activity, each group will have 10-15 minutes to present their findings, while others can provide feedback. **Closing Remarks (15 minutes)**: An organizer will convene to recap key insights gleaned from the workshop activities. Those interested in refining the artifacts created will have their information collected through a co-authorship interest form. We will emphasize that all contributions made during the workshop will be duly recognized in any resulting publications through acknowledgment if not co-authorship. Following this, there will be an opportunity for additional discussions and to formally conclude the workshop.

4 HYBRID WORKSHOP LOGISTICS

In line with our workshop's goal of incorporating diverse perspectives, we are fully dedicated to offering a hybrid experience to enhance accessibility. Before the workshop, participants will be asked to bring a device capable of accessing online platforms listed below as these tools will be essential for facilitating the hybrid experience. **Website**: The organizers will create a website for the workshop, including information about the workshop, a call for participation, an expression of interest form, the workshop schedule, and any other relevant information. **Discord**: We will create a Discord server for the workshop to facilitate participants' interactions before, during, and after the workshop. We will ask participants to introduce themselves and share why they are attending the workshop on the server. The organizers will monitor the server to foster discussions and keep participants engaged. **Zoom Video Conferencing**: All activities will be broadcast through Zoom to allow virtual participants to engage in the activities and discussions. The organizers will be on Zoom and will monitor the chat to help facilitate interactions between in-person and virtual participants. We will utilize breakout rooms to pair in-person and virtual participants to promote discussion between modalities. **Online Platforms**: To support the sharing and recording of ideas during the discussion activities, we will use Google Documents and Miro to allow workshop participants to take notes. Both of these software are available online, allowing virtual participants to access the materials.

5 CALL FOR PARTICIPANTS

We welcome 20-30 academics and practitioners who are working, researching, or interested in red teaming from fields including but not limited to CSCW, AI, HCI, sociology, communications, philosophy, psychology, and labor studies. To express interest in attending, individuals must submit a statement of interest that will include a

summary of their motivation for attending the workshop, themes they are interested in exploring, and a short biography via a Google form. The submissions will be reviewed by the workshop organizers and accepted based on the diversity of perspectives, given the focus on bringing together academic and practitioner viewpoints and approaches from diverse domains. We will advertise our workshop through social media and academic mailing lists and reach out to organizations or special interest groups that may be interested in the topic of our workshop.

6 WORKSHOP OUTCOMES

In addition to providing a venue for those already engaged with red teaming, we envision this workshop as an opportunity to highlight red teaming and its relevance to others in CSCW. Thus, the expected outcomes of this workshop include the following:

AI Red Teaming Research Network: We will bring together interdisciplinary researchers and practitioners to critically examine the current state of red teaming and how practitioners can support the humans who perform it. Through the exchange of experiences and ideas, we expect that collaborations will be formed, creating a network of researchers who will continue to work together on the topic.

Synthesis of Workshop Findings: Building on the collaborative efforts during the workshop activities, we aim to synthesize and publish key findings from the discussions and artifacts developed. We will gather insights and perspectives on AI red teaming practices by discussing developed artifacts. The resulting synthesis will offer valuable insights for practitioners and highlight avenues for further exploration by researchers.

7 ORGANIZING TEAM

Our workshop proposal includes a team of researchers experienced in critically examining labor, moderation, mental health and well-being, and RAI red-teaming.

Alice Qian Zhang is a PhD student in the Human-Computer Interaction Institute at Carnegie Mellon University. Her research explores avenues to support the human workforce with in the AI pipeline, with a particular emphasis on developing practices and tools to support mental health and well-being.

Ryland Shaw is a pre-doctoral research assistant at Microsoft Research's Social Media Collective, where he works on questions of AI ethics, technological norms, and sociotechnical systems. He has an MA in Communication from Simon Fraser University and comes from a background in documentary filmmaking.

Jacy Reese Anthis is the director of the Sentience Institute, a visiting scholar at Stanford HAI, and a PhD candidate in sociology and statistics at the University of Chicago. Jacy researches machine learning and human-AI interaction, particularly the rise of digital minds.

Ashlee Milton is a PhD candidate in computer science at the University of Minnesota, focusing on human-computer interaction. Their research investigates information retrieval systems and marginalized populations from a user perspective to support users' needs and mental well-being.

Emily Tseng is a postdoctoral researcher at Microsoft Research. Her work explores how computing technologies mediate individual, interpersonal, and structural harms, and how to create more equitable tech. Emily publishes at top-tier venues in HCI and design (CHI, CSCW), computer security and privacy (USENIX Security), and medicine (JAMA). She earned a Ph.D. in Information Science at Cornell University and a B.A. at Princeton University.

Jina Suh is a Principal Researcher in the Human Understanding and Empathy group at Microsoft Research. Her work lies at the intersection of technology and human well-being, where she examines the role of technologies, design choices, development practices, and values embedded in them in shifting power dynamics and affecting individual and organizational mental health and well-being. She received her Ph.D. in Computer Science at the University of Washington.

Lama Ahmad is a Policy Researcher at OpenAI, leading red teaming and researcher access efforts. Her work focuses on evaluating the socio-technical impact of AI systems on society. Prior to OpenAI, Lama was at Facebook, assessing the impact of social media on elections and democracy.

Ram Shankar Siva Kumar founded and leads the AI Red Team at Microsoft and co-authored *Not with a Bug, But with a Sticker: Attacks on Machine Learning Systems and What To Do About Them* [13]. He is also a Tech Policy Fellow at UC Berkeley, wherein his work on adversarial machine learning appeared notably in the National Security Commission on Artificial Intelligence (NSCAI) Final report presented to the United States Congress and the President.

Julian Posada is an Assistant Professor of American Studies at Yale University and a member of the Yale Law School's Information Society Project and the Yale Institute for Foundations of Data Science. Their research integrates theories and methods from information science, sociology, and human-computer interaction to examine how technology is developed and used within various historical, cultural, and social contexts.

Benjamin Shestakofsky is an Assistant Professor of Sociology at the University of Pennsylvania. His research centers on the relationship between work, technology, organizations, and political economy. He is the author of *Behind the Startup: How Venture Capital Shapes Work, Innovation, and Inequality* [20].

Sarah T. Roberts is an associate professor at UCLA specializing in Internet and social media policy, infrastructure, politics and culture, and the intersection of media, technology, and society. She is the faculty director of the UCLA Center for Critical Internet Inquiry (C2i2). Informed by feminist Science and Technology Studies perspectives, Roberts is keenly interested in the way power, geopolitics, and economics play out on/via the internet, reproducing, reifying, and exacerbating global inequities and social injustice.

Mary L. Gray is a Senior Principal Researcher at Microsoft Research, a Faculty Associate at Harvard University's Berkman Klein Center for Internet and Society, and a MacArthur Fellow. An anthropologist and media scholar by training, she focuses on how people's everyday uses of technologies transform labor, identity, and human rights. She maintains a faculty position in the Luddy School of Informatics, Computing, and Engineering with affiliations in Anthropology and Gender Studies at Indiana University.

REFERENCES

- [1] Hussein Abbass, Axel Bender, Svetoslav Gaidow, and Paul Whitbread. 2011. Computational red teaming: Past, present and future. *IEEE Computational Intelligence Magazine* 6, 1 (2011), 30–42.
- [2] Andrew Arshat and Daniel Etcovitch. 2018. The human cost of online content moderation. *Harvard Journal of Law and Technology* 2 (2018).
- [3] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. Philadelphia, PA.
- [4] Alistair Barr. 2015. Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms. *Wall Street Journal* (July 2015). <http://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/>
- [5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *30th Conference on Neural Information Processing Systems*. Barcelona.
- [6] Frontier Model Forum. 2024. Issue brief: What is red teaming? <https://www.frontiermodelforum.org/updates/red-teaming/>
- [7] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).
- [8] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven Connecticut.
- [9] Tarleton Gillespie. 2024. Generative AI and the Politics of Visibility. *Big Data & Society* (2024).
- [10] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (illustrated edition ed.). Harper Business, Boston.
- [11] The White House. 2023. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. <https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>
- [12] humane intelligence, SeeAI, and DEFCON AI Village. 2024.
- [13] Ram Shankar Siva Kumar and Hyrum Anderson. 2023. *Not with a Bug, But with a Sticker: Attacks on Machine Learning Systems and what to Do about Them*. John Wiley & Sons, Incorporated.
- [14] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (Jan. 2022), 34:1–34:14. <https://doi.org/10.1145/3492853>
- [15] Alexandra Olteanu, Michael Ekstrand, Carlos Castillo, and Jina Suh. 2023. Responsible AI Research Needs Impact Statements Too. *arXiv preprint arXiv:2311.11776* (2023).
- [16] Sachin R Pendse, Talie Massachi, Jalehsadat Mahdavi-moghaddam, Jenna Butler, Jina Suh, and Mary Czerwinski. 2024. Towards Inclusive Futures for Worker Wellbeing. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–32.
- [17] Sarah T. Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press. Google-Books-ID: uiCbDwAAQBAJ.
- [18] Minna Ruckenstein and Linda Lisa Maria Turunen. 2020. Re-humanizing the platform: Content moderators and the logic of care. 22, 6 (2020), 1026–1042. Publisher: Sage Publications Sage UK: London, England.
- [19] Oscar Schwartz. 2019. In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation - IEEE Spectrum. <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>
- [20] Benjamin Shestakofsky. 2017. Working algorithms: Software automation and the future of work. *Work and Occupations* 44, 4 (2017), 376–423.
- [21] Robert Soden, David Ribes, Seyram Avle, and Will Sutherland. 2021. Time for Historicism in CSCW: An Invitation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–18. <https://doi.org/10.1145/3479603>
- [22] Miriah Steiger, Timir J Bharucha, Wilfredo Torralba, Marlyn Savio, Priyanka Manchanda, and Rachel Lutz-Guevara. 2022. Effects of a Novel Resiliency Training Program for Social Media Content Moderators. In *Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Volume 4*. Springer, 283–298.
- [23] Micah Zenko. 2015. *Red Team: How to succeed by thinking like the enemy*. Basic Books.
- [24] Micah Zenko and Richard Haass. 2015. "red team: How to succeed by thinking like the enemy". <https://www.cfr.org/event/red-team-how-succeed-thinking-enemy>