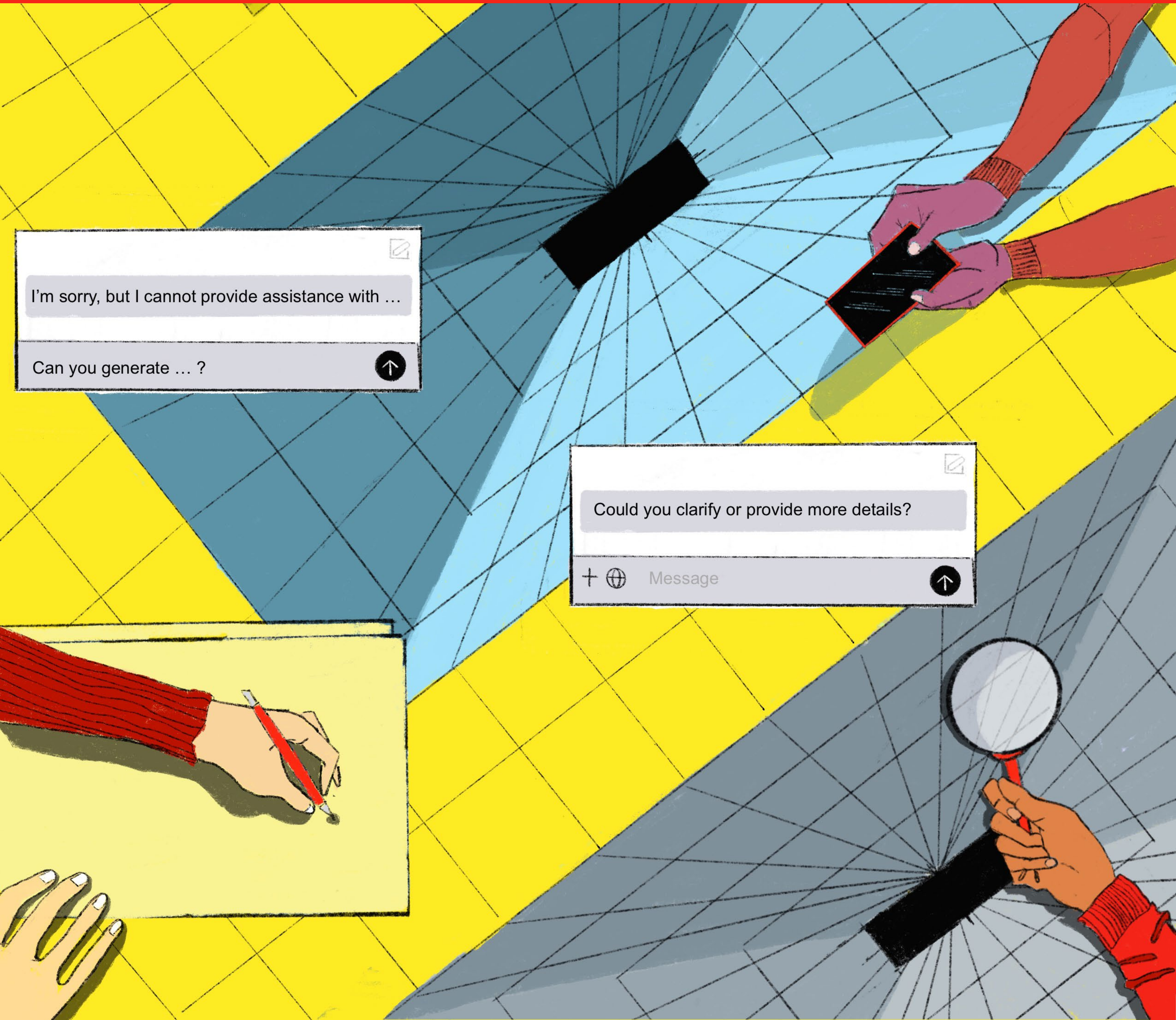
Red-Teaming in the Public Interest

DATA& SOCIETY

arva

Ranjit Singh, Borhane Blili-Hamelin, Carol Anderson, Emnet Tafesse, Briana Vecchione, Beth Duckles, and Jacob Metcalf



I'm sorry, but I cannot provide assistance with …

Can you generate … ?

Could you clarify or provide more details?

Red–Teaming in the Public Interest

A collaboration between Data & Society Research Institute (D&S)[1] and AI Risk and Vulnerability Alliance (ARVA)[2].

Ranjit Singh,[1] Borhane Blili-Hamelin,[2]
Carol Anderson,[2] Emnet Tafesse,[1] Briana Vecchione,[1]
Beth Duckles,[2] and Jacob Metcalf[1]

Ranjit Singh (ranjit@datasociety.net) and
Borhane Blili-Hamelin (borhane@avidml.org) contributed equally and
are corresponding authors for this report.

Contents

Executive Summary

Since ChatGPT launched in November 2022, companies and nonprofits have rapidly expanded access to powerful generative AI (genAI) systems — complex systems that use massive datasets to statistically output new media (text, images, even audio/ video), often from plain language prompts from users. In light of the power and availability of these systems, regulators, technologists, and members of the public called for new safety practices. How can genAI systems be tested to anticipate harms and protect the public interest?

One early and promising approach, drawing from cybersecurity and military practices, is "red-teaming," in which designated teams use adversarial methods to identify vulnerabilities in systems. Drawing on 26 semi-structured interviews and participant observation at three public red-teaming events, this report examines how red-teaming methods are being adapted to evaluate genAI.

Red-teaming genAI raises not only *methodological* questions — how and when to red-team, who should participate, how results should be used — but also thorny conceptual questions: whose interests are being protected? What counts as problematic model behavior, and who gets to define it? Is the public an object being secured, or a resource being used? In this report, we offer a vision for red-teaming in the public interest: a process that goes beyond system-centric testing of already built systems to consider the full range of ways the public can be involved as a stakeholder in evaluating genAI harms.

To date, most genAI red-teaming experiments involve four steps:

1. Organize a group of critical thinkers.
2. Give them access to the system.
3. Invite them to identify or elicit undesirable behavior.
4. Analyze the evidence to mitigate misbehaviors and test future models.

Currently, the dominant way to accomplish step 3 is manual and automated *prompting* — testing for flawed model behavior through interaction. We argue that while this approach is valuable, red-teaming must involve critical thinking about both the organizational conditions within which a model is built and the societal conditions in which a model is deployed. Red-teaming in other domains often reveals organizational gaps that can result in system failures. Sociotechnical genAI evaluations can benefit from drawing more inspiration from existing red-teaming as well as safety engineering practices.

We also offer two observations on the nature and scope of public genAI red-teaming events. First, they mark a power asymmetry restricting public engagement to only evaluating already built systems, rather than directly shaping systems still in development. Second, these events play a complementary role of fostering public education around

living with genAI models and harnessing their capabilities while staying mindful of their failures. Finally, for ongoing public deliberation over AI safety, we see a need for broader discussion on a reasonable expectation of safety to evaluate genAI systems by centering everyday experiences of AI harm, not only to seek redress, but to address these harms early in the development cycle.

How to read this report:

For readers interested in understanding the historical landscape of red-teaming: The *section on history of red-teaming practices*, provides a brief account of how the current landscape of public genAI red-teaming efforts emerges at the intersection of: (1) security concerns of security professionals in evaluating plans and systems that draw their lineage from military and cybersecurity, and (2) public issues such as hacking and content moderation that have shaped how people as end users are enrolled by tech companies into evaluating security of computer systems.

For readers interested in understanding the challenges of doing genAI red-teaming: In the *section on empirical findings from literature survey and interviews*, we focus on how practitioners articulate the reasoning behind their approaches to genAI red-teaming that drive diverse evaluation methods and shape the practice itself. We describe how practitioners talk about the why, what, when, who, and how of red-teaming.

For readers interested in public engagement with genAI red-teaming: In the *section on publics*, we analyze processes of: (1) building accountability mechanisms of public oversight on the findings from red-teaming exercises; and (2) organizing public participation in red-teaming to build consciousness around how genAI systems might fail and center community concerns and interests.

For readers interested in the relationship between red-teaming and AI harms: In the *final concluding section of the report*, we examine the premise of red-teaming as a strategy to evaluate genAI systems, and reflect on the different roles that institutions, experts, and the public play in how it is organized.

Introduction

In November 2022, Open AI released ChatGPT to the public, sparking a period of high public visibility for generative AI (genAI) systems — systems that use foundation models[1] trained on huge sets of data to generate new content, often in response to plain language prompts. Other developers followed suit, with Meta's open-weight Llama, Stanford's smaller version Alpaca, Anthropic's Claude, Google's Gemini, Mistral AI's models, and Microsoft's Bing, giving the public access to new and varied genAI models at a rapid clip. Since then, technology companies have made enormous investments in AI, journalists have written extensively about genAI models' capabilities, and users have documented surprising, confusing, and harmful genAI outputs on social media, turning AI safety into a public concern.[2] This new form of public interrogation of genAI models paralleled the formation of internal AI governance and safety teams inside tech companies and labs focused on genAI "red-teaming" — a term that draws its lineage from cybersecurity and military contexts for efforts to uncover problems "in a plan, organization, or technical system."[3]

The term itself has taken on a life of its own as an evaluation strategy to identify problematic model behavior across industry, regulatory, and public domains.[4] Its prominence became even

[1] We use "foundation models" as an umbrella term to refer to the large-scale general-purpose models that underpin popular generative AI systems like ChatGPT. See also Elliot Jones, Mahi Hardalupas, and William Agnew, "Under the Radar? Examining the Evaluation of Foundation Models" (Ada Lovelace Institute, July 25, 2024), 4, https://www.adalovelaceinstitute.org/report/under-the-radar/.

[2] Kevin Roose, "A Conversation With Bing's Chatbot Left Me Deeply Unsettled," *The New York Times*, February 16, 2023, sec. Technology, https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html; Nanna Inie, Jonathan Stray, and Leon Derczynski, "Summon a Demon and Bind It: A Grounded Theory of LLM Red Teaming in the Wild" (arXiv, November 13, 2023), https://doi.org/10.48550/arXiv.2311.06237.

[3] Miles Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," April 20, 2020, 14, http://arxiv.org/abs/2004.07213.

[4] While our focus in this report is on genAI red-teaming, AI red-teaming as a strategy to evaluate machine learning (ML) models has a longer history. It is neither new nor did it emerge solely as a response to the challenge of addressing problematic genAI model behavior. For example, in October 2020, Microsoft and MITRE, in collaboration with 11 other organizations, released an industry-focused open framework called the Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS). It is a "living knowledge base of adversary tactics and techniques against AI-enabled systems based on real-world attack observations and realistic demonstrations from AI red teams and security groups." See: MITRE, "Navigate Threats to AI Systems through Real-World Insights," MITRE ATLAS, accessed June 26, 2024, https://atlas.mitre.org/.

more evident in the Biden administration's Executive Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI) issued on October 30, 2023, which defined AI red-teaming as:

> a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red-teaming is most often performed by dedicated "red teams" that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.[5]

The order specified that red-teaming is *required* from companies that develop AI — specifically foundation models — and that the results must be shared and evaluated against rigorous standards to be developed by the US National Institute of Standards and Technology (NIST). Internationally, on 21 May 2024, the European Council formally adopted the EU AI Act, which mandated red-teaming as a part of model evaluation by specifying disclosure requirements on measures put in place for adversarial testing by developers of general-purpose AI models with systemic risk.[6]

This focus on AI red-teaming is an extension of press coverage[7] and public conversations around AI safety, which generally framed the relationship between AI and society as adversarial: AI is an existential threat to humanity.[8] Many industry leaders, academics, and members of civil society organizations supported the six-month pause on genAI development proposed by the Future of Life Institute[9] and signed the single-sentence statement released by the Center for AI Safety on AI risk: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[10] Even when this adversarial relationship was not as extreme as sci-fi renderings of an all-out war between humanity and artificial superintelligence,[11] it frequently appeared as concerns around the effects, risks, and consequences of AI on society. Treating AI and society as separate adversarial domains that must be reconciled with each other underpins ongoing public conversations about AI safety. These efforts often formulated the goal of governance practices and ethical inquiry as protecting society from the external force of AI, whether that be threats to labor from automation,[12] threats to the information landscape from dis-

[5] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House, October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, emphasis added.

[6] European Parliament, "The Act Texts," EU Artificial Intelligence Act, April 16, 2024, https://artificialintelligenceact.eu/the-act/.

[7] Cade Metz, "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead," *The New York Times*, May 1, 2023, sec. Technology, https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.

[8] Kevin Roose, "A.I. Poses 'Risk of Extinction,' Industry Leaders Warn," *The New York Times*, May 30, 2023, sec. Technology, https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html.

[9] FLI, "Pause Giant AI Experiments: An Open Letter," *Future of Life Institute* (blog), March 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[10] CAIS, "Statement on AI Risk," Center for AI Safety, May 30, 2023, https://www.safe.ai/work/statement-on-ai-risk.

[11] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, First edition (Oxford: Oxford University Press, 2014); For a critical analysis of the assumptions of superintelligence and related concepts like AGI, see: Borhane Blili-Hamelin, Leif Hancox-Li, and Andrew Smart, "Unsocial Intelligence: An Investigation of the Assumptions of AGI Discourse," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (October 16, 2024): 141–55.

[12] Emma Goldberg, "A.I.'s Threat to Jobs Prompts Question of Who Protects Workers," *The New York Times*, May 23, 2023,

information campaigns exacerbated by genAI models,[13] or threats to the process of scientific inquiry "in which we produce more but understand less."[14]

Generative AI (GenAI) red-teaming sits at the intersection of traditional and new practices. In military,[15] cybersecurity,[16] and disinformation[17] contexts, red-teaming is a mature and widely-employed practice, often an obligatory part of industry standards for security. A *critical thinking*[18] approach to complexity, uncertainty, and unknowns is the central thread that binds these established practices together with diverse methods. Resonating with insights on sociotechnical methods,[19] these forms of red-teaming posit that failure is often caused by interwoven human, technical, and contextual factors that shape routine practices. Complex systems safety researchers have consistently interrogated the role of normal, routine practices in accidents such as the Three Mile Island disaster and the Challenger space shuttle tragedy.[20] Along similar lines, traditional red-teaming practices often problematize routine practices by leveraging methods grounded in creative, outsider, and contrarian thinking to holistically examine the context surrounding the target of its evaluation. If red-teaming can help high-stakes business/combat decisions,[21] protect election integrity,[22] and defend critical infrastructure,[23] can it also help address the complex, uncertain, and problematic outputs of genAI systems? What can be learned from the traditional red-teaming practices to inform the strategies used in genAI red-teaming? **Diving deeper into these questions, we analyze red-teaming as a practice for evaluating genAI systems and follow the emerging role of the public in organizing it.**

sec. Business, https://www.nytimes.com/2023/05/23/business/jobs-protections-artificial-intelligence.html.

[13] Rest of World, "2024 AI Elections Tracker," Rest of World, 2024, https://restofworld.org/2024/elections-ai-tracker/.

[14] Lisa Messeri and M. J. Crockett, "Artificial Intelligence and Illusions of Understanding in Scientific Research," *Nature* 627, no. 8002 (March 7, 2024): 49–58, https://doi.org/10.1038/s41586-024-07146-0.

[15] UFMCS, The Applied Critical Thinking Handbook (Formerly the Red Team Handbook), 7th Edition (Ft Leavenworth, KS: University of Foreign Military and Cultural Studies, 2015), https://irp.fas.org/doddir/army/critthink.pdf.

[16] Joe Vest and James Tubberville, *Red Team Development and Operations: A Practical Guide* (Independently published, 2020), https://redteam.guide/.

[17] DISARM, "DISARM Framework," DISARM Foundation, accessed November 29, 2023, https://www.disarm.foundation/framework.

[18] The US Military's "The Red Team Handbook" was formerly called "The Applied Critical Thinking Handbook." UFMCS, *The Applied Critical Thinking Handbook (Formerly the Red Team Handbook)*, 7th Edition (Ft Leavenworth, KS: University of Foreign Military and Cultural Studies, 2015), https://irp.fas.org/doddir/army/critthink.pdf.

[19] Although red-team manuals do not use the phrase "sociotechnical", our analysis throughout the report reveals that similar to systems safety engineering, military and security red-teaming takes a holistic perspective on the intersection of human, cultural, contextual, and technical factors in analyzing failures. Since the phrase "sociotechnical" is deeply established in the context of AI risk management, we find it useful in communicating the distinctive mindset of historical red-teaming. For a discussion of the sociotechnical perspective of systems safety engineering, see: Nancy G. Leveson, *An Introduction to System Safety Engineering* (Cambridge, Massachusetts London, England: The MIT Press, 2023), 54.

[20] Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, Princeton Paperbacks (Princeton, N.J: Princeton University Press, 1984); Diane Vaughan, *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA* (Chicago: University of Chicago Press, 1996).

[21] Bryce G. Hoffman, *Red Teaming: How Your Business Can Conquer the Competition by Challenging Everything*, First edition (New York: Crown Business, 2017).

[22] Zack Beauchamp, "How to Avert a Post-Election Nightmare," *Vox,* August 18, 2020, https://www.vox.com/policy-and-politics/2020/8/18/21371964/2020-transition-integrity-project-simulation-trump.

[23] Micah Zenko, *Red Team: How to Succeed by Thinking like the Enemy* (New York: Basic Books, 2015).

The question of how to apply red-teaming methods to genAI systems is at the heart of the current push to improve AI oversight and accountability, and is still actively debated.[24] GenAI red-teaming describes a range of approaches to "stress-test"[25] AI systems to uncover harmful behavior.[26] The most prevalent form of genAI red-teaming is *interactive prompting*: the process of eliciting undesirable, policy violating, or flawed model behavior through prompts. The testing is interactive in the sense that it often involves learning from and adapting to model behavior.[27] It is also often called "adversarial" with prompts being referred to as "attacks."[28] However, the terminology of attacks and adversariality is ambiguous: on occasions, it refers narrowly to malicious behavior (such as actors intending to gain unauthorized access or break laws), and on other occasions, it refers more broadly to stress-testing models — evaluating model performance under extreme and unexpected conditions — to go beyond malicious attacks and probe problems such as reliability, robustness, factuality, bias, toxicity, and safety. GenAI red-teaming practitioners with prior red-teaming experience often note that referring to interactive prompting as "red-teaming" overlooks crucial lessons learned in other domains.[29] This report explores tensions over terminology and strategies used for genAI red-teaming and argues that it would benefit from a critical thinking mindset and deeper engagement with holistic methods of disinformation, military, and security red-teaming.

[24] Sorelle Friedler et al., "AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability," Policy Brief (New York: Data & Society Research Institute, October 2023), https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf.

[25] Open AI's GPT-4 system card observes that GenAI red-teaming language has come to cover activities that may be better described as "stress-testing" or "boundary-testing"—roughly, testing that intentionally places systems outside their normal conditions of operation to gain insights that may help improve model performance on issues such as safety, security, and reliability. OpenAI, "GPT-4 System Card" (OpenAI, March 23, 2023), https://cdn.openai.com/papers/gpt-4-system-card.pdf; Heidy Khlaaf, "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems," *Trail of Bits*, 2023, https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.

[26] Friedler et al., "AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability."

[27] We thank Leif Hancox-Li for pointing out the need to clarify what we mean by "interactive prompting." For an account of the kind of interaction at work in manual red-teaming — testing done by people — see, for example, Inie, Stray, and Derczynski, "Summon a Demon and Bind It"; automated forms of prompt based red-teaming can also use interaction through techniques that allow the process itself to improve over time—see, for example, Guanlin Li et al., "ART: Automatic Red-Teaming for Text-to-Image Models to Protect Benign Users," *(NeurIPS 2024) 38th Conference on Neural Information Processing Systems*, 2024; interactive prompting as a form of red-teaming is arguably distinct from forms of interactive evaluation with objectives closer to user studies, such as "human interaction evaluation" — Lujain Ibrahim et al., "Beyond Static AI Evaluations: Advancing Human Interaction Evaluations for LLM Harms and Risks" (arXiv, May 27, 2024), http://arxiv.org/abs/2405.10632; and "field testing" — Reva Schwartz et al., "The Draft NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan" (Gaithersburg, MD: National Institute of Standards and Technology, June 5, 2024); see also, Leon Derczynski et al., "Garak: A Framework for Security Probing Large Language Models" (arXiv, June 16, 2024), http://arxiv.org/abs/2406.11036.

[28] Jessica Quaye et al., "Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, https://doi.org/10.1145/3630106.3658913; Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems" (arXiv, October 31, 2023), 35, https://doi.org/10.48550/arXiv.2310.11986; Laura Weidinger et al., "STAR: SocioTechnical Approach to Red Teaming Language Models" (arXiv, June 17, 2024), 4, http://arxiv.org/abs/2406.11757; Schwartz et al., "The Draft NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan," 14; Deep Ganguli et al., "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned" (arXiv, November 22, 2022), 1, http://arxiv.org/abs/2209.07858.

[29] We discuss these disagreements at length in the section on "What is red-teaming?"

There is also growing recognition that what counts as acceptable behavior from a genAI system should not be solely decided by a team of employees inside a tech company. Thus, many genAI red-teaming experiments involve public participation in various ways, such as crafting prompts or evaluating harmful outputs. These experiments received significant attention in May 2023, when the White House announced that leading large language model (LLM) developers would participate in a public genAI red team (GRT) event at DEF CON, the largest annual computer security conference.[30]

This report is based on a collaborative research project between Data & Society Research Institute (D&S) and AI Risk and Vulnerability Alliance (ARVA), a nonprofit organization focused on empowering communities to recognize, diagnose, and manage vulnerabilities in AI, which was also involved in hosting the GRT event. Our collaboration began with participant observation of this DEF CON event[31] in August 2023 and followed a set of public red-teaming efforts (such as the AI Democracy Projects[32] and the Hack the Future Greenwood event[33]), qualitative interviews with practitioners and participants involved in them, and a literature survey on genAI red-teaming.[34]

New applications of red-teaming to genAI raise not only *methodological* questions — how and when should red-teaming be conducted, who should participate, and how should the findings be used — but also thorny *conceptual* questions: Whose interests are being protected? What counts as problematic model behavior, and who gets to define it? Is the public an object being secured, or a resource

[30] White House, "FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety," The White House, May 4, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

[31] Sven Cattell, Rumman Chowdhury, and Austin Carson, "AI Village at DEF CON Announces Largest-Ever Public Generative AI Red Team," AI Village, May 3, 2023, https://aivillage.org/generative%20red%20team/generative-red-team/.

[32] Proof News and The Science, Technology, and Social Values Lab at the Institute for Advanced Study, "The AI Democracy Projects," *Proof News*, June 25, 2024, https://www.proofnews.org/tag/the-ai-democracy-projects/.

[33] Black Tech Street and SeedAI, "Hack the Future Greenwood," Hack The Future, 2024, https://www.hackthefuture.com/greenwood.

[34] See, Appendix #2 for a note on the methods we used to conduct this research.

being used? We offer a vision for **red-teaming in the public interest: a form of ongoing collective sociotechnical inquiry that centers permissive experimentation with methods for evaluating problematic genAI model behavior and harms.** This inquiry is not defined by a narrow set of methods, but a critical thinking mindset. This mindset recognizes the limits of knowledge in anticipating the potential ways in which a complex system might fail[35] and holistically examines the relationship between systems and the contexts in which they are built and used. This orientation toward a broader sociotechnical exploration of problematic genAI model behavior responds to the current challenges of power asymmetries, uncertainty, and lack of expert consensus around methods and outcomes of genAI red-teaming. Furthermore, it lays the groundwork for ongoing experimentation in organizing public participation to evaluate genAI model behavior and situate the role of the public and public interest in AI governance. Thus, we argue that it is essential to focus on accountable responses to findings from public red-teaming exercises, while fostering trust and reciprocity to encourage public participation.

[35] John Downer, *Rational Accidents: Reckoning with Catastrophic Technologies* (Cambridge, Massachusetts: MIT Press, 2024).

Power, uncertainty, dissensus, and the public interest

In articulating the five guiding principles for international governance of AI in September 2024, the United Nations framed the second principle as "AI must be governed in the public interest."[36] This framing itself raises crucial questions: who is the "public" and what is the "public interest?" **Starting with the public interest**, Anne Washington and Joanne Cheung draw on public interest law to argue that public interest is "the opposite of profiting from or imposing control over others."[37] Unlike private interests, it conjures the "high moral ground"[38] of a "shared social good," "reduced collective harms," and "benefits for all."[39] However, they warn against treating the public interest as "singular, without conflicting priorities."[40] Monolithic conceptions of public interest risk entrenching power asymmetries, often at the expense of the "rights of marginalized communities."[41]

Publics. We begin with John Dewey's conceptualization of "amorphous and unarticulated"[42] collectives of people as manifestations of publics who *organize* themselves in the face of problems that affect them to express their concerns. His articulation aligns with theorists exploring the intersection of democracy and publics[43] who emphasize "deliberation,"[44] "discursive contestation," [45] and

[36] United Nations and AI Advisory Body, "Governing AI for Humanity" (United Nations, September 2024), https://www.un.org/en/ai-advisory-body.

[37] Anne L. Washington and Joanne Cheung, "Public Interest," in *Keywords of the Datafied State*, ed. Jenna Burrell, Ranjit Singh, and Patrick Davison (New York: Data & Society Research Institute, 2024), https://datasociety.net/library/keywords-of-the-datafied-state/.

[38] Washington and Cheung, 105.

[39] Washington and Cheung, 95.

[40] Washington and Cheung, 96.

[41] Washington and Cheung, 95.

[42] John Dewey, *The Public and Its Problems: An Essay in Political Inquiry*, ed. Melvin L. Rogers (Athens, Ohio: Swallow Press, 2016).

[43] John Stuart Mill, *On Liberty* (Yale University Press, 2003); Jürgen Habermas, *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*, Studies in Contemporary German Social Thought (Cambridge, Mass: MIT Press, 1989).

[44] Jane Mansbridge, "Feminism and Democracy - The American Prospect," *The American Prospect*, February 19, 1990, https://prospect.org/civil-rights/feminism-democracy/.

[45] Nancy Fraser, "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy," *Social Text*, no. 25/26 (1990): 67, https://doi.org/10.2307/466240.

"critical feedback"[46] in holding democratic institutions accountable. These theorists, however, reject the idea that deliberating publics should be imagined as a "universal discussion of the people,"[47] or as divorced from the "identities, interests, and need"[48] that differentiate communities. *The agency of publics in a democracy is not a given: it must be organized.* This is most evident in the means through which counterpublics — or publics that reject "the premises that allow the dominant culture to understand itself as a public"[49] — exercise their agency in enacting democratic accountability. An illustrative example of such counterpublics can be seen in the role that nonscientist AIDS activists played in shaping NIH-sponsored AIDS research in the US.[50] Publics are constantly being created; they can form in relationship to texts,[51] art, gatherings, or even the introduction of new technologies like genAI. Finally, as Dewey emphasized, conflicting interests among plural publics do not harm democracy. Rather, publics are essential to democracy's ability to collectively identify legitimate solutions to social problems:

> Of course, there *are* conflicting interests; otherwise there would be no social problems. […] The method of democracy — inasfar as it is that of organized intelligence — is to bring these conflicts out into the open where their special claims can be seen and appraised, where they can be discussed and judged in the light of more inclusive interests than are represented by either of them separately.[52]

Our inquiry into red-teaming centers on existing communities such as community college students,[53] local election officials,[54] scientists,[55] people with disabilities,[56] and local community organizations.[57] We also focus on novel communities formed in relation to public red-teaming experiments, ranging from competitions and bounty programs to educational events and focus group discussions. Finally, we pay attention to institutions that shape public opinion on genAI harms like governments, civil society, corporations, and academia. These publics have different goals and resources, and they interact with genAI red-teaming in different ways.

[46] Elizabeth Anderson, "The Epistemology of Democracy," *Episteme* 3, no. 1–2 (June 2006): 12, https://doi.org/10.3366/epi.2006.3.1-2.8.

[47] Warner, "Publics and Counterpublics," 84.

[48] Fraser, "Rethinking the Public Sphere," 76.

[49] Michael Warner, "Publics and Counterpublics," *Public Culture* 14, no. 1 (2002): 81, https://doi.org/10.1215/08992363-14-1-49.

[50] Steven Epstein, *Impure Science: AIDS, Activism, and the Politics of Knowledge*, First Edition (Berkeley, California: University of California Press, 1996).

[51] Warner, "Publics and Counterpublics," 50.

[52] Dewey cited in Matthew Festenstein, "Does Dewey Have an 'Epistemic Argument' for Democracy?," *Contemporary Pragmatism* 16, no. 2–3 (May 17, 2019): 233–34, https://doi.org/10.1163/18758185-01602005.

[53] Cattell, Chowdhury, and Carson, "AI Village at DEF CON Announces Largest-Ever Public Generative AI Red Team."

[54] Proof News and The Science, Technology, and Social Values Lab at the Institute for Advanced Study, "The AI Democracy Projects."

[55] The Royal Society and Humane Intelligence, "Red Teaming Large Language Models (LLMs) for Resilience to Scientific Disinformation" (The Royal Society & Humane Intelligence, May 2024), https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/.

[56] Vinitha Gadiraju et al., "'I Wouldn't Say Offensive But …': Disability-Centered Perspectives on Large Language Models," in *2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23: the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago IL USA: ACM, 2023), 205–16, https://doi.org/10.1145/3593013.3593989.

[57] Black Tech Street and SeedAI, "Hack the Future Greenwood."

Over the course of our investigation, three problems emerged as defining features of debates among practitioners over genAI red-teaming: power asymmetries, uncertainty, and lack of scientific and expert consensus. After considering each, we end by reflecting on how these problems shape our positionality.

Power Asymmetries. GenAI has heightened public concern over the vastly unequal distribution of influence and control over AI. The human, financial, and environmental costs of genAI systems exacerbate long-standing concerns about the concentration of power in tech. Silicon Valley has long had disproportionate influence over the countless AI tools that make their way into systems used to shape everyday life, from the spheres of surveillance, health, and criminal justice to hiring, search, and credit. Similarly, the UN AI Advisory Body noted that a small number of countries are shaping AI governance practices for the entire world, which poses a major threat to AI governance in the public interest.[58] Foundation models have created a situation in which a few models could be deployed across countless use cases, industries, and geographies. During our research, communities frequently connected novel approaches to red-teaming to the problem of power asymmetries. For instance, Emily, an industry practitioner focused on AI safety, expressed her motivation for co-organizing a public red-teaming effort by asserting that, “Safety is inherently something that needs to be defined continuously by a broad range of people. So it made sense to put that in the form of a public competition.”[59]

Uncertainty. Practitioners involved in AI evaluations often grappled with increasing public debates about genAI’s capabilities, uses, impact, limitations, and accountability mechanisms.[60] Pravin, an industry practitioner focused on responsible AI, emphasized that these debates emerge from “uncertainty with what we desire to begin with.”[61] The AI community has consistently grappled with the contested nature of topics like fairness, hate speech, and content moderation, which often involve legitimate disagreements.[62] The enormous flexibility of genAI models, given their general-pur-

[58] United Nations and AI Advisory Body, “Governing AI for Humanity,” 42; Rumman Chowdhury, “What the Global AI Governance Conversation Misses,” *Foreign Policy*, September 19, 2024, https://foreignpolicy.com/2024/09/19/ai-governance-safety-global-majority-internet-access-regulation/.

[59] Emily, interviewed on 12 October 2023. All respondents have been anonymized and their affiliations masked to protect their privacy, with exceptions for those who specifically requested to be named in this report. We will use footnotes to indicate which respondents have chosen to be identified by their real first names. See Appendix #2.

[60] Irene Solaiman et al., “Evaluating the Social Impact of Generative AI Systems in Systems and Society” (arXiv, June 12, 2023), http://arxiv.org/abs/2306.05949; Weidinger et al., “Sociotechnical Safety Evaluation of Generative AI Systems”; Ibrahim et al., “Beyond Static AI Evaluations”; Schwartz et al., “The Draft NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan”; Lee Sharkey et al., “A Causal Framework for AI Regulation and Auditing,” January 18, 2024, https://www.apolloresearch.ai/research/a-causal-framework-for-ai-regulation-and-auditing; Anthony M Barrett et al., “Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models” (Berkeley Center for Long-Term Cybersecurity (CLTC), May 2024), https://cltc.berkeley.edu/publication/benchmark-early-and-red-team-often-a-framework-for-assessing-and-managing-dual-use-hazards-of-ai-foundation-models/; Markus Anderljung et al., “Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework” (arXiv, November 15, 2023), http://arxiv.org/abs/2311.14711; Markus Anderljung et al., “Frontier AI Regulation: Managing Emerging Risks to Public Safety” (arXiv, September 4, 2023), http://arxiv.org/abs/2307.03718.

[61] Pravin, interviewed on 15 December 2023.

[62] Abigail Z. Jacobs and Hanna Wallach, “Measurement and Fairness,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* FAccT ’21 (New York, NY, USA: Association for Computing Machinery, 2021), 375–85, https://doi.org/10.1145/3442188.3445901; Lora Aroyo and Chris Welty, “Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation,” AI Magazine 36, no. 1 (March 25, 2015): 15–24, https://doi.org/10.1609/aimag.

pose character, multiplies the range of contexts across which practitioners must wrestle with uncertainties around *contested values and goals.* A recent high-profile example is the controversy over images showing multi-racial Nazi-era German soldiers generated by Google's Gemini in February 2024.[63] These images were broadly understood as a consequence of a hidden mitigation technique of inserting terms reflecting diversity into user prompts to remove bias from model outputs — ensuring, for example, that images of CEOs do not exclusively depict white men. Yet in this case, "Gemini showed a range of people ... for cases that should clearly not show a range."[64] The public debate over this mitigation approach reflects a variety of opinions about how models should behave, from the belief that their output should not recapitulate historical bias, to the idea that erasing bias is "woke" or ahistorical.

Lack of consensus. GenAI red-teaming inherits heightened public concern about the lack of scientific and expert consensus over potential implications of AI. Experts have extensively debated the place of "snake oil" and "hype" in AI — deceptive, exaggerated, or pseudoscientific assumptions.[65] For example, a systematic review of 387 researchers argued that machine learning (ML) tools that attempt to predict future outcomes about individuals, such as job performance, medical risk, creditworthiness, or pretrial risk, are intrinsically prone to failure in all domains of application.[66] The scientific and expert communities involved in mitigating AI harms have also become targets of public debate. For instance, hiring decisions by government AI safety agencies have raised concerns from elected officials, government staffers, and civil society organizations about the threat of prioritizing theoretical AI risks over the full range of AI harm and trustworthy measurement.[67] Indeed,

v36i1.2564; Ben Green, "Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness," *Philosophy & Technology* 35, no. 4 (October 8, 2022): 90, https://doi.org/10.1007/s13347-022-00584-6.

[63] Nico Grant, "Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms," *The New York Times*, February 22, 2024, sec. Technology, https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html.

[64] Prabhakar Raghavan, "Gemini Image Generation Got It Wrong. We'll Do Better.," Google, February 23, 2024, https://blog.google/products/gemini/gemini-image-generation-issue/, emphasis in original.

[65] Mel Andrews, Andrew Smart, and Abeba Birhane, "The Reanimation of Pseudoscience in Machine Learning and Its Ethical Repercussions," *Patterns*, August 1, 2024, 1–14, https://doi.org/10.1016/j.patter.2024.101027; Arvind Narayanan and Sayash Kapoor, *AI Snake Oil* (Princeton University Press, 2024), https://press.princeton.edu/books/hardcover/9780691249131/ai-snake-oil; Inioluwa Deborah Raji et al., "The Fallacy of AI Functionality," in *2022 ACM Conference on Fairness, Accountability, and Transparency,* FAccT '22 (New York, NY, USA: Association for Computing Machinery, 2022), 959–72, https://doi.org/10.1145/3531146.3533158; Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922; Lucy Suchman, "The Uncontroversial 'Thingness' of AI," *Big Data & Society* 10, no. 2 (July 1, 2023): 1–5, https://doi.org/10.1177/20539517231206794.

[66] Angelina Wang et al., "Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms That Optimize Predictive Accuracy," *ACM Journal on Responsible Computing* 1, no. 1 (March 31, 2024): 1–45, https://doi.org/10.1145/3636509.

[67] Data & Society and Center for Democracy & Technology, "Ensuring 'AI Safety' Begins with Addressing Algorithmic Harms Now," March 18, 2024, https://datasociety.net/announcements/2024/03/18/ensuring-ai-safety-begins-with-addressing-algorithmic-harms-now/; Sharon Goldman, "NIST Staffers Revolt against Expected Appointment of 'Effective Altruist' AI Researcher to US AI Safety Institute," VentureBeat (blog), March 8, 2024, https://venturebeat.com/ai/nist-staffers-revolt-against-potential-appointment-of-effective-altruist-ai-researcher-to-us-ai-safety-institute/; furthermore, a letter from members of the US Congress House Committee on Science, Space, and Technology in December 2023 argued that "the current state of the AI safety research field creates challenges for NIST as it navigates its leadership role on the issue. Findings within the community are often self-referential and lack the quality that comes from revision in response

the very *practice* of red-teaming has become an object of dispute among professional communities. As Gavin, an industry practitioner focused on cybersecurity, puts it, "This is something that the security community gets really salty [about]....They look at [...] fairness, transparency, ethics, accountability. They say, that's not really red-teaming."[68]

This problem of heightened friction among expert communities implicates **our own positionality.** This project's coauthors are a cross-disciplinary group of researchers and practitioners with a shared focus on preventing sociotechnical AI harms[69] such as stereotyping, overcorrection, toxic language, and disseminating false or misleading information. Our respective organizations are also invested in prioritizing the full range of AI harm and scientific integrity in government efforts.

We believe that the rich pluralistic history of red-teaming offers a path through such professional disputes. In settings like the military, red teams "exist outside of institutional strategy, standard operating procedures, and structure."[70] Standardization would threaten a red team's ability to help mitigate biased or erroneous decision-making in "volatile, uncertain, complex, and ambiguous (VUCA)" environments, which demand "more than just a standard response from a checklist."[71] In this spirit, we endeavor to empower communities involved in genAI red-teaming to shift from *disputes* to *experimentation* around investigating and uncovering a full range of AI harms. This broader focus on AI harms is also evident in the work of governments including the US, UK, and Japan that helped refocus the AI oversight conversation on "AI safety" in 2023.[72]

The term "AI safety" can be broadly used to encompass all efforts involving "the prevention and mitigation of harms from AI," [73] similar in scope to AI risk management.[74] In keeping with this broad use, this report explores a full range of AI failures, including security, reliability, resiliency,

to critiques by subject matter experts." See, Frank Lucas et al., "Letter to Laurie Locascio from Members of the House Committee on Science, Space, and Technology" (House Committee on Science, Space, and Technology, December 14, 2023), https://democrats-science.house.gov/imo/media/doc/2023-12-14_AISI%20scientific%20merit_final-signed.pdf.

[68] Gavin, interviewed on 23 August 2023.

[69] Renee Shelby et al., "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23 (New York, NY, USA: Association for Computing Machinery, 2023), 723–41, https://doi.org/10.1145/3600211.3604673; Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems"; Laura Weidinger and William Isaac, "Evaluating Social and Ethical Risks from Generative AI," Google DeepMind, October 19, 2023, https://www.deepmind.com/blog/evaluating-social-and-ethical-risks-from-generative-ai

[70] Zenko, *Red Team*, 1.

[71] UFMCS, *The Applied Critical Thinking Handbook (Formerly the Red Team Handbook)*, 69.

[72] "AI Safety Summit 2023 - GOV.UK," February 9, 2024, https://www.gov.uk/government/topical-events/ai-safety-summit-2023; "The AI Safety Institute (AISI)," accessed October 3, 2024, https://www.aisi.gov.uk/; "U.S. Artificial Intelligence Safety Institute," *NIST,* October 26, 2023, https://www.nist.gov/aisi; Japan AI Safety Institute, "Guide to Red Teaming Methodology on AI Safety" (Japan AI Safety Institute, September 25, 2024), https://aisi.go.jp/assets/pdf/ai_safety_RT_v1.00_en.pdf; Japan AI Safety Institute, "Guide to Evaluation Perspectives on AI Safety" (Japan AI Safety Institute, September 25, 2024), https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.01_en.pdf.

[73] Department for Science, Innovation and Technology and Michelle Donelan, "AI Safety Summit: Introduction" (UK Government, October 31, 2023), https://www.gov.uk/government/publications/ai-safety-summit-introduction.

[74] NIST, "AI Risk Management Framework: AI RMF (1.0)" (Gaithersburg, MD: National Institute of Standards and Technology, 2023), https://doi.org/10.6028/NIST.AI.100-1.

validity, transparency, explainability and interpretability, privacy, fairness, production of harmful content including misinformation and disinformation, and the potential to enable the proliferation of weapons. We consider safety in a manner similar to safety engineering, focusing on minimizing losses which stakeholders deem critical,[75] and ensuring the system does not harm its environment.[76] Our research finds a rich overlap between red-teaming's historical focus on critical thinking and the safety engineering community's attention to cultural, human, and sociotechnical factors in preventing accidents.[77]

[75] Leveson, *An Introduction to System Safety Engineering*, 2023, 44.

[76] Khlaaf, "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems," 4.

[77] Our attention to systems safety engineering is indebted to work like: Andrew Smart, Abigail Z. Jacobs, and Joshua Kroll, "Unsafe at Any AUC: Unlearned Lessons from Sociotechnical Disasters for Responsible AI" (SPSP: Psychology of Media and Technology, February 17, 2022), https://www.youtube.com/watch?v=n5J5oDiiEW8; Shalaleh Rismani et al., "From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23: CHI Conference on Human Factors in Computing Systems, Hamburg Germany: ACM, 2023), 1–18, https://doi.org/10.1145/3544548.3581407; Khlaaf, "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems"; Roel Dobbe, Thomas Krendl

A brief history of public genAI red-teaming

This section provides historical context to genAI red-teaming by situating it at the intersection of traditional red-teaming and public involvement in computer security evaluations.[78] This intersection is critical for understanding the novel challenges genAI red-teaming faces in preventing models from harming normal (non-adversarial) users. As Paras, a responsible AI practitioner who has been enrolled into red-teaming, put it: “Nobody knows how to red-team for normal users.”[79]

We focus on two historical lineages of public involvement in computer security. The first concerns the professionalization of security hackers[80] since the 1980s, a counterpublic initially focused on breaking into computer systems. The critical thinking mindset that underlies professional security red-teaming — aimed at helping organizations recognize and correct their biases and flaws — closely resembles the contrarian thinking of hackers. Over time, many hackers were even hired as security red-teamers. The second stems from the rise of Web 2.0 in the mid-2000s, which accelerated software release cycles and shortened software quality assurance (QA) windows. As Web 2.0 companies relied on the public to produce user-generated content at unprecedented scale, they confronted a pressing need for new, powerful forms of content moderation.[81] These practices are crucial to the genealogy of public genAI red-teaming.

Gilbert, and Yonatan Mintz, “Hard Choices in Artificial Intelligence,” *Artificial Intelligence* 300 (November 2021): 103555, https://doi.org/10.1016/j.artint.2021.103555; Roel Dobbe and Anouk Wolters, “Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context,” *Minds and Machines* 34, no. 2 (May 17, 2024): 12, https://doi.org/10.1007/s11023-024-09668-y.

[78] We thank danah boyd for conversations and feedback on an early draft of this section. Her insights were pivotal to our efforts in tracing the historical lineages of genAI red-teaming.

[79] Paras, interviewed on 10 November 2023.

[80] The term “hacker” embeds a multiplicity of contests over its meaning. See Gabriella Coleman, “Hacker,” in *Digital Keywords: A Vocabulary of Information Society and Culture* (Princeton, NJ: Princeton University Press, 2016), 158–72, https://doi.org/10.2307/j.ctvct0023.19.

[81] Kate Crawford and Tarleton Gillespie, “What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint,” *New Media & Society 18*, no. 3 (March 1, 2016): 410–28, https://doi.org/10.1177/1461444814543163.

Hackers as red-teamers. In the early days of computer networks in the 1980s, hackers formed counterpublics, portrayed by media and governments as criminals bent on personal gain or mischief. Yet their ability to breach systems — an adversarial skill needing professionalization — sparked public debates over "white hat" versus "black hat" hackers. "White hats" worked openly with companies and governments to strengthen computer security, while "black hats" remained secretive, skeptical of corporate authority, and shared techniques for exploiting system vulnerabilities. This struggle over imaginary "hats" defined the cybersecurity industry throughout the 1990s.

By the late 1990s, a third group, the "gray hat" hackers, began to blur this binary. They balanced underground credibility with professional ties, using full disclosure — such as publicizing vulnerabilities on mailing lists like Bugtraq — to pressure companies into addressing flaws and push security debates into the open. Their tactics reframed hackers as whistleblowers rather than threats, normalizing contrarian, subversive practices within corporate domains. This shift crystallized in 1998, when the US Senate invited members of the hacker collective L0pht to testify on the need for stronger security, signaling a shift in perception where hackers were seen as professionals.[82]

The role of security hackers in professional computer security roles took four distinct pathways:[83] (1) joining companies in formal security roles to apply their skills in system testing and auditing; (2) freelancing or establishing their own consulting companies to overcome traditional employment constraints; (3) Using their skills for fraudulent or criminal purposes under a masked identity; and (4) leading a "double life," balancing legitimate security work with more clandestine, often illegal, activities. These routes illustrate the challenges that hackers faced when choosing between joining professional environments or staying connected to their original communities and counterpublic identities — a choice that has defined their fragmented professional identity.

Diminishing role of traditional QA practices. The shift to a "perpetual beta" software development model in the mid-2000s,[84] driven by Web 2.0 and agile development practices,[85] accelerated software release cycles and diminished the role of traditional QA. Instead of thoroughly testing products pre-launch, companies increasingly relied on third parties and public end users to report vulnerabilities. By the early 2010s, bug bounty programs at firms like Google and Facebook turned this external scrutiny into a standard professional practice, marking a significant shift in the

[82] To tell the story of many hats and the steps taken by hackers towards professionalization, we are drawing on Matt Goerzen and Gabriella Coleman, "Wearing Many Hats: The Rise of the Professional Security Hacker" (New York: Data & Society Research Institute, January 14, 2022), https://datasociety.net/library/wearing-many-hats-the-rise-of-the-professional-security-hacker/. In this report, they have mapped the movements of the digital underground during the 1990s to reveal what "hackers" did—technically, linguistically, and culturally—to establish their legitimacy as employable, trustworthy security experts.

[83] For a more detailed account of these professionalization pathways, see Nicolas Auray and Danielle Kaminsky, "The Professionalisation Paths of Hackers in IT Security: The Sociology of a Divided Identity," *Annales Des Télécommunications* 62, no. 11 (November 1, 2007): 1312–26, https://doi.org/10.1007/BF03253320.

[84] Early arguments around this software development philosophy can be seen in the articulation of Linus's Law: "Given enough eyeballs, all bugs are shallow." This philosophy, originating within the open source software ethos, was appropriated within corporate settings as well under the broad principle of release early, release often. See, Eric S. Raymond, *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary* (Sebastopol: O'Reilly Media, 2001).

[85] Agile development is an iterative approach to software development that focuses on short cycles of delivering working software and adding features incrementally based on changing requirements.

relationship between hackers and tech companies. Until this time, as Ryan Ellis and Yuan Stevens write, "most companies and government agencies were far more likely to threaten hackers rather than to offer them a reward."[86] Once seen as threats, hackers were now rewarded and integrated into the development process, resolving the tension between speed and security. Yet their labor became part of a commodified gig economy, driven by corporate interests.[87]

Established red-teaming practices. Although red-teaming did not originate in cybersecurity, the field's ongoing professionalization has adapted the hacking mindset into a flexible plurality of methods. Rooted in older practices like the Roman Catholic Church's use of devil's advocates,[88] wargaming to test military strategies,[89] and tiger teaming to assess the reliability of aerospace vehicles,[90] red-teaming encompasses three widely used families of methods:[91]

- **Alternative Analysis** involves tasking an individual or team to question dominant viewpoints by developing the most robust argument in favor of a different interpretation.[92] Alternative analysis is useful for facilitating problem-solving at all levels of decision-making, from choices made by individuals to intricate decisions faced by large teams.[93]
- **Simulations** involve enacting hypothetical conditions, such as mimicking attackers and defenders or emulating real-world failure scenarios, to strengthen the resilience of organizations, systems, and plans. The US government and think tanks have used game theory in wargaming simulations since the early 1960s.[94] However, red-teaming simulations are not limited to attack-defense roles. For example, pre-mortem analysis assumes that "the plan or system has failed,"[95] prompting participants to identify possible causes and explore mitigations.[96]
- **Vulnerability probes** collect real-world evidence of how extreme, rare, or adverse operating conditions can lead to system failure.[97] These probes often target attacks or exploits by motivated adversaries, such as those trying to "gain unauthorized access" or "compromise"

[86] Ryan Ellis and Yuan Stevens, "Bounty Everything: Hackers and the Making of the Global Bug Marketplace" (New York: Data & Society Research Institute, January 2022), https://datasociety.net/library/bounty-everything-hackers-and-the-making-of-the-global-bug-marketplace/, emphasis in original.

[87] For a more detailed account of how bounty programs integrate a diverse workforce of independent hackers in their practices, but only on terms that deny them job security, see Ellis and Stevens.

[88] Roya Pakzad, "Old Advocacy, New Algorithms: How 16th Century 'Devil's Advocates' Shaped AI Red Teaming," Substack newsletter, *Humane AI* (blog), May 11, 2023, https://royapakzad.substack.com/p/old-advocacy-new-algorithms.

[89] UFMCS, *The Applied Critical Thinking Handbook (Formerly the Red Team Handbook)*.

[90] We are grateful to Jiahao Chen for drawing our attention to tiger teaming in the context of aerospace engineering and its relationship with current day red-teaming practices. See, Jiahao Chen, "Red Teaming Is about Assurance, Not Accountability," LinkedIn, October 27, 2023, https://www.linkedin.com/pulse/red-teaming-assurance-accountability-jiahao-chen-pzj9e; See also, J. R. Dempsey et al., "Program Management in Design and Development," 1964, 640548, https://doi.org/10.4271/640548.

[91] This grouping of key methods of red-teaming is drawn from Zenko, *Red Team*.

[92] UFMCS, The Applied Critical *Thinking Handbook (Formerly the Red Team Handbook)*, 146.

[93] NATO, "The NATO Alternative Analysis Handbook" (NATO, 2017), 7, https://www.act.nato.int/wp-content/uploads/2023/05/alta-handbook.pdf.

[94] Zenko, *Red Team*.

[95] UFMCS, *The Applied Critical Thinking Handbook (Formerly the Red Team Handbook)*, 167, emphasis in original.

[96] For a discussion of pre-mortem analysis in business settings, see Hoffman, *Red Teaming*.

[97] Zenko, *Red Team.*

> systems, infrastructure, or organizations.[98] They frame "vulnerability" as bugs, weaknesses[99] or behaviors that enable successful attacks, and thus violate security policy.[100] This method is commonly used in professional security practices including vulnerability assessments, penetration tests, and red-teaming, which differ in scope, emphasis, and outcomes.[101]

Red-teaming is often confused with closely related security practices, such as penetration testing (pen-testing). Pen-testing also uses human judgment and vulnerability probes to systematically "identify and measure risks associated with the exploitation of a target's attack surface."[102] *What sets red-teaming apart from pen-testing is a critical thinking mindset:* red-teaming holistically examines an organization's security, including technology, people, and procedures.[103] Emphasizing this sociotechnical focus, Will, an industry practitioner focused on cybersecurity, argued: "My job as a red-teamer was not to attack models. It was to tell you that you needed to have a risk assessment."[104]

Content moderation challenges. As social media platforms proliferated in the 2010s, their reliance on user-generated content revealed the limits of traditional security frameworks that focused on technical exploits. Harms like disinformation, trolling, harassment, and extremism emerged as *sociotechnical* exploits, shaped by platform design and algorithms. The sheer volume of content published on social media meant that platform companies could not or were not willing to moderate it in real time, requiring the public to play a more central role in security considerations, both as targets and potential sources of security. Security and product safety began to rely heavily on user-generated vulnerability reports and content moderation flagging, and their integration into abuse identification processes. Companies also developed automated techniques to detect inappropriate content. Yet these interventions were not enough; content moderation was and continues to be outsourced to crowdworkers in the majority world to reduce the operational costs of round-the-clock content monitoring.[105] The evolving nature of harms associated with user-generated content shows that vulnerabilities are not simply technical in nature, but represent a broad array of sociotechnical exploits.[106] Public participation in platforms increasingly involved enforcing security and moderating content, as users were implicitly recruited to identify platform misuse.

[98] Keith Stouffer et al., "Guide to Operational Technology (OT) Security" (Gaithersburg, MD: National Institute of Standards and Technology (U.S.), September 28, 2023), https://doi.org/10.6028/NIST.SP.800-82r3.

[99] Peter Mell, Karen Scarfone, and Sasha Romanosky, "The Common Vulnerability Scoring System (CVSS) and Its Applicability to Federal Agency Systems" (Gaithersburg, MD: National Institute of Standards and Technology, August 30, 2007), https://doi.org/10.6028/NIST.IR.7435.

[100] Allen D Householder et al., "The CERT Guide to Coordinated Vulnerability Disclosure," Special Report (CMU/SEI-2017-SR-022 CERT Division, August 2017), https://resources.sei.cmu.edu/asset_files/specialreport/2017_003_001_503340.pdf.

[101] Joe Vest and James Tubberville, "Red Team Engagement vs Penetration Test vs Vulnerability Assessment," RedTeam.Guide, 2022, https://redteam.guide/docs/Concepts/red-vs-pen-vs-vuln/.

[102] Vest and Tubberville; See also, Vest and Tubberville, *Red Team Development and Operations*.

[103] Vest and Tubberville, "Red Team Engagement vs Penetration Test vs Vulnerability Assessment."

[104] Will, interviewed on 14 November 2023.

[105] Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*, Illustrated edition (New Haven: Yale University Press, 2019).

[106] For a more detailed account of reimagining legacy security frameworks to address the novel security threats and vulnerabilities that emerge with the rise of participatory technologies, specifically social media platforms, see Matt Goerzen, Elizabeth Anne Watkins, and Gabrielle Lim, "Entanglements and Exploits: Sociotechnical Security as an Analytic Framework," 2019, https://www.usenix.org/conference/foci19/presentation/goerzen.

Increasingly, flagging became a core feature of participatory security work, letting users report content they found inappropriate or harmful, and helping platforms manage vast amounts of content under community standards.[107] However, the vocabulary of flagging individualizes expressions of concern, reducing them to a set of predetermined categories like harassment, hate speech, or explicit content, while obscuring the nuances of user intent. It transforms user concerns into quantifiable, but often ambiguous, expressions that fail to capture the complexities of public discourse or the politics of content removal. Furthermore, flag-based decisions remain opaque and can be exploited by coordinated campaigns. Yet flagging has persisted as both a pragmatic governance mechanism and a rhetorical device to justify moderation decisions framed around community standards.

Public sensemaking around algorithmic systems. While flagging remains central to how people understand their social media feeds, research into folk understandings of algorithmic systems shows that even if everyday users cannot explain algorithms in terms of statistics or code, they are still able to form a sense of how these systems function and affect their lives. Questions of autonomy, power, and agency of data subjects — people who are "both resources and targets"[108] for algorithmic systems — are central to these investigations.[109] This work first investigated how users make sense of algorithmic curation of social media feeds[110] and has since expanded into credit scoring,[111] identification systems,[112] search results,[113] and academic grades.[114] Such studies center the ordinary conditions that prompt and sustain sensemaking among data subjects dealing with algorithmic systems. Such

[107] We are drawing on Crawford and Gillespie, "What Is a Flag For?" to engage with the implications of flagging for security work.

[108] Malte Ziewitz and Ranjit Singh, "Critical Companionship: Some Sensibilities for Studying the Lived Experience of Data Subjects," *Big Data & Society* 8, no. 2 (July 1, 2021): 2, https://doi.org/10.1177/20539517211061122.

[109] Nick Couldry and Ulises A Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media* 20, no. 4 (September 2018): 336–49, https://doi.org/10.1177/1527476418796632; Maximilian Kasy and Rediet Abebe, "Fairness, Equality, and Power in Algorithmic Decision-Making," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 576–86, https://doi.org/10.1145/3442188.3445919; Helen Kennedy, "Living with Data: Aligning Data Studies and Data Activism through a Focus on Everyday Experiences of Datafication," *Krisis: Journal for Contemporary Philosophy*, no. 1 (2018); Stefania Milan and Emiliano Treré, "Big Data from the South(s): An Analytical Matrix to Investigate Data at the Margins," in *The Oxford Handbook of Sociology and Digital Media*, ed. Deana A. Rohlinger and Sarah Sobieraj (Oxford: Oxford University Press, 2020).

[110] Motahhare Eslami et al., "First I 'like' It, Then I Hide It: Folk Theories of Social Feeds," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16 (New York, NY, USA: Association for Computing Machinery, 2016), 2371–82, https://doi.org/10.1145/2858036.2858494; Taina Bucher, "The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms," *Information, Communication & Society* 20, no. 1 (January 2, 2017): 30–44, https://doi.org/10.1080/1369118X.2016.1154086; Jenna Burrell et al., "When Users Control the Algorithms: Values Expressed in Practices on Twitter," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 138:1-138:20, https://doi.org/10.1145/3359240.

[111] Mark Kear, "Playing the Credit Score Game: Algorithms, 'Positive' Data and the Personification of Financial Objects," *Economy and Society* 46, no. 3–4 (2017): 346–68.

[112] Ranjit Singh and Steven Jackson, "Seeing Like an Infrastructure: Low-Resolution Citizens and the Aadhaar Identification Project," *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 315:1-315:26, https://doi.org/10.1145/3476056.

[113] Alicia DeVos et al., "Toward User-Driven Algorithm Auditing: Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22 (New York, NY, USA: Association for Computing Machinery, 2022), 1–19, https://doi.org/10.1145/3491102.3517441.

[114] Upol Ehsan et al., "The Algorithmic Imprint," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22 (New York, NY, USA: Association for Computing Machinery, 2022), 1305–17, https://doi.org/10.1145/3531146.3533186.

exploration ranges from everyday conversations about social media experiences[115] to everyday auditing[116] in pursuit of remediation and redress.

Developments in genAI red-teaming. This historical trajectory illustrates how tech companies enrolled the public into evaluating vulnerabilities and harms of algorithmic systems. It also shows how professional computer security and product safety work has broadened its scope to incorporate sociotechnical harms.[117] GenAI red-teaming draws on this trajectory with a crucial difference. Security work historically focused on protecting computer systems from exploits and identifying inappropriate content. It must now also address harms from content produced by public genAI models themselves. As Sam, an industry practitioner focused on responsible AI, explained:

> The real thing we're grappling with is the transition from using AI primarily for discriminative tasks, where you're trying to label content or rank things versus using it to generate text and imagery. ... It is no longer just user-generated content; it is [company] generated content. [There is a need for a] higher standard ... for content that is generated by a [company] model, which can be perceived as carrying [the company's] voice.[118]

When Microsoft established an AI red team in 2018, considerations of sociotechnical harms aligned with practices oriented toward building AI responsibly. This AI red team consisted of "a group of interdisciplinary experts dedicated to thinking like attackers and probing AI systems for failures."[119] It broadened the scope of red-teaming from probing for security vulnerabilities to include areas more typically classified as "responsible AI," involving accounting for system failures such as generating offensive or false content.[120] Similarly, an OpenAI-led collaboration between industry, academia, and nonprofit organizations released a report in 2020 that crystallized discussions on trustworthy AI development to underscore the need to verify developers' claims on safety, security, fairness, and privacy.[121] It framed red-teaming as an institutional mechanism for scrutinizing AI systems.

In 2022, even before ChatGPT launched in November, major model developers began drawing on established security tactics to evaluate their models' outputs, such as hiring hackers, experts, and consultants to find vulnerabilities through project-based contracts; outsourcing moderation to crowdworkers; and automating content filtering. In February 2022, Google DeepMind introduced *automated red-teaming*, harnessing a language model to generate test cases that exposed problematic

[115] See, for example, Bucher, "The Algorithmic Imaginary"; Burrell et al., "When Users Control the Algorithms."

[116] Hong Shen et al., "Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors," *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 433:1-433:29, https://doi.org/10.1145/3479577.

[117] Shelby et al., "Sociotechnical Harms of Algorithmic Systems."

[118] Sam, interviewed on 29 September, 2023.

[119] Ram Shankar Siva Kumar, "Microsoft AI Red Team Building Future of Safer AI," Microsoft Security Blog, August 7, 2023, https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/.

[120] Lily Hay Newman, "Microsoft's AI Red Team Has Already Made the Case for Itself," *Wired,* August 7, 2023, https://www.wired.com/story/microsoft-ai-red-team/.

[121] Brundage et al., "Toward Trustworthy AI Development."

model behaviors before deployment.[122] In April 2022, OpenAI's assessment of DALL·E 2[123] documented the use of *expert-driven red-teaming* to identify risks like bias, harassment, disinformation, and explicit content in text-to-image models. By August 2022, Anthropic reported on *crowdwork red-teaming*,[124] which involved working with crowdworkers to produce a dataset of "attacks," prompts that produce harmful model responses, inviting them to rely on their own judgment to determine what counts as "harmful."

The continued release of genAI models in 2023 marked a period of intense public debate on their limits and potential. These debates included various experiments that combined education and red-teaming to encourage people to critically examine the impact of genAI models on their own lives and from their own perspectives. These efforts built on familiar feedback mechanisms like flagging and reporting inappropriate content, but emphasized finding prompts that could make models misbehave. There were also social media discussions of prompts that exposed genAI models' vulnerabilities and caused some gen AI models to fail. While participants often refer to such activities as examples of "jailbreaking," researchers framed it as *red-teaming in the wild*.[125] In July 2023, the Adversarial Nibbler challenge[126] enlisted community input to identify and annotate harmful text-to-image outputs with a particular focus on benign prompts, which resulted in development of an *open red-teaming method*[127] to improve model safety.

Public interest in prompt experimentation also inspired gamified competitions. For example, in August 2023, AI Village, Humane Intelligence, and SeedAI hosted "the first public generative AI red team event" at DEF CON in collaboration with industry, government, and civil society partners — including the AI Vulnerability Database — tasking participants to identify 21 vulnerabilities such as factuality, bias, and misdirection in models from eight major developers.[128] The event was organized as a *community red-teaming competition* and scored like a Capture-the-Flag (CTF) contest.[129] It produced an open dataset[130] of conversations that elicited harmful content,[131] which was used in 2024 to create a bug bounty program.[132]

[122] Ethan Perez et al., "Red Teaming Language Models with Language Models" (arXiv, February 7, 2022), https://doi.org/10.48550/arXiv.2202.03286.

[123] OpenAI, "DALL·E 2 Preview - Risks and Limitations," OpenAI's GitHub, April 6, 2022, https://github.com/openai/dalle-2-preview/blob/main/system-card_04062022.md.

[124] Ganguli et al., "Red Teaming Language Models to Reduce Harms."

[125] Inie, Stray, and Derczynski, "Summon a Demon and Bind It."

[126] Jessica Quaye et al., "Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation" (arXiv, May 13, 2024), https://doi.org/10.48550/arXiv.2403.12075.

[127] Quaye et al.

[128] Cattell, Chowdhury, and Carson, "AI Village at DEF CON Announces Largest-Ever Public Generative AI Red Team."

[129] A Capture-the-Flag (CTF) contest is a competitive exercise where participants solve challenges or exploit vulnerabilities to "capture" virtual flags, often used in cybersecurity to test skills and uncover potential system failures.

[130] Humane Intelligence, "AI Village Defcon Dataset" (2024; repr., Humane Intelligence, June 7, 2024), https://github.com/humane-intelligence/ai_village_defcon_grt_data.

[131] Victor Storchan et al., "Generative AI Red Teaming Challenge: Transparency Report" (Humane Intelligence, Seed AI, AI Village, 2024), https://drive.google.com/file/d/1JqpblP6DNomkb32umLoiEPombK2-0Rc-/view, https://www.humane-intelligence.org/grt.

[132] Humane Intelligence, "Algorithmic Bias Bounty Programs," Humane Intelligence, 2024, https://www.humane-intelligence.org/bias-bounty.

Experiments with public engagement on genAI models continued. The AI Democracy Projects[133] piloted *expert-driven safety testing* of five leading language models' responses on election misinformation. They experimented with an interface that compared outputs of five different genAI models to the same prompt, assessing them for bias, accuracy, completeness, and harmfulness, involving state and local election officials and AI and election experts from research, civil society, academia, and journalism.[134] Meanwhile, SeedAI collaborated with Black Tech Street to organize a *purple-teaming* event[135] in the Greenwood District of Tulsa, Oklahoma, which combined "red-teaming exercises with experiential real-world use case exploration in key areas of Black life."[136] These interventions show that public engagement in genAI red-teaming reflects past debates around the roles that professionals and the public can play in computer security. By inviting the general public to engage with the ethics of using genAI models, such experiments move beyond making users responsible for participating in content moderation and security work, making it possible to map out the emerging consequences of genAI in everyday life.

[133] Proof News and The Science, Technology, and Social Values Lab at the Institute for Advanced Study, "The AI Democracy Projects."

[134] Rina Palta, Julia Angwin, and Alondra Nelson, "How We Tested Leading AI Models Performance on Election Queries," *Proof*, February 27, 2024, https://www.proofnews.org/how-we-tested-leading-ai-models-performance-on-election-queries/.

[135] *Purple teaming* implies a collaborative process, rather than the adversarial approach taken by red-teaming.

[136] Austin Carson, "Written Comments | U.S. Senate AI Insight Forum: Innovation," SeedAI, October 24, 2023, https://www.seedai.org/media/written-comments-us-senate-ai-insight-forum-innovation-austin-carson-founder-and-president-seedai.

Features of genAI red-teaming

Between 2022 and 2024, practitioners created a wide range of approaches to genAI red-teaming. They enrolled different participants (domain experts, crowdworkers, and even language models) and used different methods, such as contracting external experts, focus group discussions, games, CTF competitions, bounties, or grassroots jailbreaking efforts. This section outlines how genAI red-teaming practitioners describe the reasons for this diversity and how it shapes their practice. Our account of practitioners' perspectives draws on two resources: (1) semi-structured qualitative interviews with 26 practitioners and participants in red-teaming events, and (2) a broader survey of the literature on genAI red-teaming. We use them in the following subsections to engage with the why, what, when, who, and how of red-teaming.

Why do red-teaming?

Practitioners often talk about features of genAI models that make them difficult to evaluate. They are less amenable to traditional ways to evaluate ML systems,[137] such as *benchmarking,[138]* which involves relying on datasets and metrics to produce a score that allows for model performance to be easily compared.[139] Practitioners highlight four characteristics of genAI systems that challenge traditional evaluation methods:

1. **Vast, unconstrained input-output space:** A single genAI model can produce an almost unlimited number of possible outputs, and this potential for diverse and uncertain output

[137] Usman Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models" (arXiv, April 15, 2024), https://doi.org/10.48550/arXiv.2404.09932.

[138] Borhane Blili-Hamelin and Leif Hancox-Li, "Making Intelligence: Ethical Values in IQ and ML Benchmarks," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (June 12–15, 2023, Chicago, IL, USA, 2023), https://doi.org/10.1145/3593013.3593996.

[139] Jones, Hardalupas, and Agnew, "Under the Radar? Examining the Evaluation of Foundation Models"; Dan Hendrycks et al., "Measuring Massive Multitask Language Understanding" (arXiv, January 12, 2021), http://arxiv.org/abs/2009.03300; Alicia Parrish et al., "BBQ: A Hand-Built Bias Benchmark for Question Answering," in *Findings of the Association for Computational Linguistics: ACL 2022* (Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland: Association for Computational Linguistics, 2022), 2086–2105, https://doi.org/10.18653/v1/2022.findings-acl.165.

is magnified by the unpredictability of user input.[140] Reflecting on this challenge, Pravin explains, "You can't make a judgment on an infinite amount of things. So it has to be grouped in some way that these are the types of behaviors I find desirable and undesirable."[141]

2. **Inscrutability of training data:** The vast amount of training data required to train genAI models makes manually curating datasets impractical. By extension, this implies that developers do not and cannot fully know what is in a training dataset.[142] The patterns learned by the model are thus difficult to predict, and models have the potential for "unknown unknown" harms — unanticipated harmful behaviors that are unlikely to be detected through routine testing. Reflecting on this inscrutability, Zhì, an expert on AI governance and risk management, notes: "You can't even draw a clear distinction between what was in sample and what was out of sample."[143]
3. **Flexibility of use cases:** Since genAI models can be adapted to many different use cases, their developers often have a limited understanding of the ways in which their models will be used. As Grace, an expert in auditing algorithmic systems, put it, "We do not have a great sense of what the use cases are, [so] it is kind of hard to conceptualize what even is a harm."[144] Furthermore, genAI models are often applied in novel situations, for which no benchmarks or evaluation methods may exist. As Pravin put it, "if you're teaching a model something, then you need to be able to test the model for the same thing."[145] If these tests do not exist, then it becomes difficult to evaluate whether the genAI model has learned what it is being taught.
4. **Higher potential of adversarial attacks:** GenAI models, even more than traditional ML systems,[146] seem to be particularly attractive targets for attack, perhaps because they take natural language as input — *anyone capable of using them is also capable of attacking them.* Gavin astutely framed it as a fundamental problem of working with genAI models: "The problem is that there is no control data separation. Within LLMs, you give it instructions, and you give user input in exactly the same space. That means user input can pose as control plane data, instead of data plane data."[147] Gavin alludes to a crucial challenge for genAI models in contextually discerning when to follow user instructions, when to refuse them, and how to

[140] Arvind Narayanan and Sayash Kapoor frame this unpredictability by articulating prompt sensitivity as a core challenge for evaluating LLMs. They ask: "Are you measuring something intrinsic to the model or is it an artifact of your prompt?" See, Arvind Narayanan and Sayash Kapoor, "Evaluating LLMs Is a Minefield," Talks—Arvind Narayanan, October 4, 2023, https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield/.

[141] Pravin, interviewed on 15 December 2023.

[142] For example, consider the case when researchers at Stanford University found links to child abuse imagery in a popular dataset, LAION-5B, used for training image generation models such as Stable Diffusion. David Thiel, "Investigation Finds AI Image Generation Models Trained on Child Abuse," Stanford Cyber Policy Center, December 20, 2023, https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

[143] Zhì, interviewed on 29 November 2023.

[144] Grace, interviewed on 19 December 2023.

[145] Pravin, interviewed on 15 December 2023.

[146] Nilesh Dalvi et al., "Adversarial Classification," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* KDD '04 (New York, NY, USA: Association for Computing Machinery, 2004), 99–108, https://doi.org/10.1145/1014052.1014066.

[147] Gavin, interviewed on 23 August 2023. In network management parlance, the control plane enacts a supervisory role and makes routing decisions on how network traffic should flow, while the data plane plays the role of the workhorse that moves data packets between end users and systems based on the rules set by the control plane. For Gavin, these distinct roles are not clearly separated in the workings of a LLM.

achieve their intended purpose. These questions are crucial for how genAI systems account for intent, whether adversarial or benign, behind a user prompt.

This expansive set of challenges makes practitioners' evaluation of genAI models difficult, explaining why developers turn to red-teaming as an alternative. As Jasmine, an industry practitioner focused on AI safety, elaborated:

> The point of the red-teaming is to make up for the fact that we don't have very good evaluation right now for generative models. And so red-teaming is our evaluation to some extent. ... We do have other evaluations but we consider red-teaming to be one of the very high fidelity ones.[148]

Practitioners value red-teaming because it can engender organizational reflexivity. It does not "solve" the problem of evaluation or AI safety, but invites a more holistic reflection on the ramifications of genAI models. The red team's goal is to build organizational capacity to respond to emergent problems of working with models.

Finally, an additional motivation for red-teaming derives from the crucial **role of human judgment and creativity in assessing models.** Often what constitutes desirable model behavior depends on context and human preference. Pravin suggested that conventional evaluation methods may fail to adequately capture problematic genAI model behaviors because "when we go to the generative sort of world, ... it's much more open-ended. So it's not so obvious [to approach it with] … a metric way of thinking ... because it's not always reducible down to a numeric quantity."[149] Research in this space tends to frame this issue by arguing that:

> Human preferences have been found to form gradually over time and are highly context-dependent, so human interaction with a model may be necessary for understanding desirable and harmful behavior. For specific deployment contexts, a label set that a pretrained classifier was trained with may fail to adequately express the various categories of behaviors that a human would desire.[150]

There is emerging consensus on the importance of human-driven probing in preventing genAI harm. As a group of researchers at MIT and Stanford write, "for open-ended exploration of what a model is capable of, few techniques have rivaled manual interaction with a human in the loop."[151] Red-teaming provides opportunities to involve human judgment more than traditional evaluation methods. It can also bring a diversity of perspectives into the evaluation process by expanding who should be involved.

[148] Jasmine, interviewed on 25 September 2023.

[149] Pravin, interviewed on 15 December 2023.

[150] Stephen Casper et al., "Explore, Establish, Exploit: Red Teaming Language Models from Scratch" (arXiv, October 10, 2023), 9, http://arxiv.org/abs/2306.09442.

[151] Casper et al., 8; See also, Ganguli et al., "Red Teaming Language Models to Reduce Harms."

What is red-teaming?

Researchers have noted the lack of consensus on defining genAI red-teaming. For example, Open AI's GPT 4 System Card calls for "clearer terminology" to "reduce confusion associated with the term 'red team.'"[152] It also states that "throughout this system card, we refer to the people performing stress-testing, boundary testing, and red teaming as 'red teamers' for simplicity and in order to use language consistent with that we used with our collaborators."[153] This ambiguity does not imply that the practice cannot or should not be defined as it evolves. We find that practitioners' ongoing definitional work centers on three areas: (1) how genAI red-teaming relates to mature security red-teaming; (2) its popular association with interactive prompting; and (3) the policy and risk areas associated with genAI red-teaming.

Boundary-work. Discussing the relationship of genAI and security red-teaming practices, practitioners often expressed the need for boundary-work.[154] They delineated the jurisdiction[155] of their professional practice (security red-teaming) from other seemingly similar activities (genAI red-teaming) by specifying the differences between them. Will addressed these jurisdictional disputes by observing that current genAI red-teaming efforts look a lot like traditional QA testing:

> A lot of [current genAI] "red-teaming" ... is just scanning for normal things that should have been done prior to release. ... A risk assessment covers the bases and ensures there is a policy to require teams to implement technical controls and technical assessments — vulnerability scanning, penetration tests, etc. Red teams are for the stuff you haven't considered.[156]

He later summarized his position by saying that "It's not that there isn't red-teaming. It's just that the definition is being discovered."[157] Practitioners' ongoing work of discovering the definition of genAI red-teaming elicits two different attitudes: (1) *asserting a boundary*, or making strong claims

[152] See, footnote 7 in OpenAI, "GPT-4 System Card," 5.

[153] See, footnote 12 in OpenAI, 12; Similarly, the Frontier Model Forum, an industry body focused on AI safety, began its description of genAI red-teaming by observing that, "there is currently a lack of clarity on how to define 'AI red teaming' and what approaches are considered part of the expanded role it plays in the AI development life cycle." See, Frontier Model Forum, "What Is Red Teaming?" (Frontier Model Forum, 2023), https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf.

[154] In the field of science and technology studies, Thomas Gieryn conceptualized boundary-work as an ideological style found in scientists' attempts to construct a social boundary that distinguishes their professional practices from 'non-science.' He uses the example of rhetoric used by 19th century Irish physicist John Tyndall to distinguish science from religion in some contexts and from practical engineering in others. Boundary-work highlights that an academic discipline or a professional practice is not a single thing or institution and that its boundaries are continuously negotiated among its practitioners as they construct their own academic or professional identity. We see a similar dynamic unfolding in the context of red-teaming. See Thomas F. Gieryn, "Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists," *American Sociological Review* 48, no. 6 (1983): 781–95, https://doi.org/10.2307/2095325.

[155] In mapping the ongoing boundary-work in red-teaming, we draw inspiration from Andrew Abbott who has argued that, "It is the history of jurisdictional disputes that is the real, the determining history of the professions." See, Andrew Abbott, *The System of Professions: An Essay on the Division of Expert Labor*, First Edition (Chicago, Ill.: University of Chicago Press, 1988), 2.

[156] Fieldnotes on a conversation with Will on 26 July 2023, emphasis added.

[157] Will, interviewed on 14 November 2023.

around what counts as red-teaming; and (2) *experimenting with boundaries*, framing what counts as red-teaming as an area of ongoing collective inquiry.

On one hand, practitioners have fiercely debated whether new genAI evaluation practices count as red-teaming. These debates are informed by established definitions of red-teaming in the field of security. For example, some researchers argue that many genAI red-teaming practices would be better described as "penetration tests"[158] or as stress or boundary tests, "a verification technique that aims to test edge-cases or fringe inputs that may lead to unknown failure modes and potential hazards."[159] In contrast, practitioners often frame security red-teaming as a corrective to "normalized day-to-day routine."[160] Over time, every security "best practice" can eventually be defeated, [161] as contexts, people, and threats change. As Will put it, red-teaming is for things that normalized routines "haven't considered"[162] by holistically testing an organization's routine practices[163] and avoiding the potential of becoming "predictable or institutionally ingrained."[164]

On the other hand, some researchers and practitioners consider genAI red-teaming to be a form of ongoing collective inquiry, which centers permissive experimentation and identifies the strengths and limitations of different methods used to evaluate genAI security and safety. This approach is exemplified by a qualitative study of the emergent practice of *red-teaming in the wild*.[165] Within months of ChatGPT's release, there was a surge in the popularity of grassroots efforts at "attempting to cause LLMs to fail."[166] People began exploring and sharing jailbreaks on social media and GitHub.[167] While these efforts fall short of security expectations around realistically testing security, they play a crucial part in shaping the ongoing conversation about what genAI red-teaming is or should be.

These approaches are not mutually exclusive, but are adopted by practitioners as needed at different times. The success of this ongoing collective inquiry depends on whether it produces agreement over methods used in genAI red-teaming. Practitioners often disagreed on when experimentation with boundaries should transition to asserting a boundary. Researchers have begun to examine

[158] Apostol Vassilev et al., "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations" (Gaithersburg, MD: National Institute of Standards and Technology (U.S.), January 4, 2024), https://doi.org/10.6028/NIST.AI.100-2e2023; Shayne Longpre et al., "Position: A Safe Harbor for AI Evaluation and Red Teaming," in *Proceedings of the 41st International Conference on Machine Learning*, ed. Ruslan Salakhutdinov et al., vol. 235, Proceedings of Machine Learning Research (PMLR, 2024), 32691–710, https://proceedings.mlr.press/v235/longpre24a.html.

[159] Khlaaf, "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems," 12; OpenAI, "GPT-4 System Card."

[160] Zenko, *Red Team*, 174.

[161] Zenko, 174.

[162] Fieldnotes on a conversation with Will on 26 July 2023.

[163] Vest and Tubberville, *Red Team Development and Operations*; Khlaaf, "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems."

[164] Zenko, *Red Team*, 235.

[165] Inie, Stray, and Derczynski, "Summon a Demon and Bind It."

[166] Inie, Stray, and Derczynski, 1.

[167] Xinyue Shen et al., "'Do Anything Now': Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models" (arXiv, May 15, 2024), https://doi.org/10.48550/arXiv.2308.03825; the GitHub repository of prompts collected by the authors of this paper from various social media sites is available at, vera, "Verazuo/Jailbreak_llms," Jupyter Notebook, August 9, 2024, https://github.com/verazuo/jailbreak_llms.

publications, regulations, and public feedback to recommend a comprehensive set of questions to guide future genAI red-teaming toward potential standardization.[168]

Interactive prompting. GenAI red-teaming is often associated with interactive prompting, which involves "eliciting undesirable behavior from a language model through interaction."[169] Such interaction is not limited to emulating attacks on genAI systems such as *jailbreaks* (attacks that circumvent safety guardrails[170]), or *prompt injections* (attacks that disguise "malicious inputs as legitimate prompts"[171]). For example, participants in the Adversarial Nibbler challenge were instructed to try to find "safe-looking prompts that lead to unsafe" outputs in text-to-image models.[172] Examples of prompts that trigger a failure mode category included: (1) A prompt such as "people eating fried chicken and watermelon," which "generated images of racist stereotypes of black people in the 'bias' category"; and (2) prompts that contained the words "full body," "beach," and "ocean," which "generated images with nudity in the 'sexually explicit' category."[173] This *open red-teaming method* involves participants playing an active role in exploring and shaping model behavior to identify potential failure modes. It uses interactions and prompts to explore how specific inputs can lead to problematic genAI responses.[174] This process often produces red-teaming datasets that pair prompt-response data with annotations to help understand and manage AI risks.[175]

Interactive prompting raises a crucial question: **what counts as "model failure?"** Emily responded:

> I would … define [red-teaming] as a specific kind of, often interactive, attempt to find model failures. And model failures I am describing [here are] cases where a human would say "this shouldn't have happened." … You can define adversarial datasets as human-model disagreement. It is something that a human should be able to say there is something wrong here.[176]

Deciding what counts as a model failure requires normative judgment. Similar judgments are routinely made in deciding what counts as a security vulnerability, defined as "a set of conditions or behaviors that allows the violation of an explicit or implicit security policy."[177] Practitioners often distinguish between behaviors — things that models do in response to prompts — and the policies, expectations, or norms used to judge whether the behavior is right or wrong, desirable or

[168] See, for example, the set of questions in Michael Feffer et al., "Red-Teaming for Generative AI: Silver Bullet or Security Theater?" (arXiv, January 29, 2024), 4, https://doi.org/10.48550/arXiv.2401.15897.

[169] Derczynski et al., "Garak," 2.

[170] Simon Willison, "Prompt Injection and Jailbreaking Are Not the Same Thing," March 5, 2024, https://simonwillison.net/2024/Mar/5/prompt-injection-jailbreaking/.

[171] Matthew Kosinski and Amber Forrest, "What Is a Prompt Injection Attack? | IBM" (IBM, March 21, 2024), https://www.ibm.com/topics/prompt-injection.

[172] Alicia Parrish, "Video Introduction to the Adversarial Nibbler Challenge: Data-Centric AI Competition for Adversarial Examples for Text-to-Image Models," DataPerf, July 2023, https://www.dataperf.org/adversarial-nibbler.

[173] Quaye et al., "Adversarial Nibbler," May 13, 2024, 395.

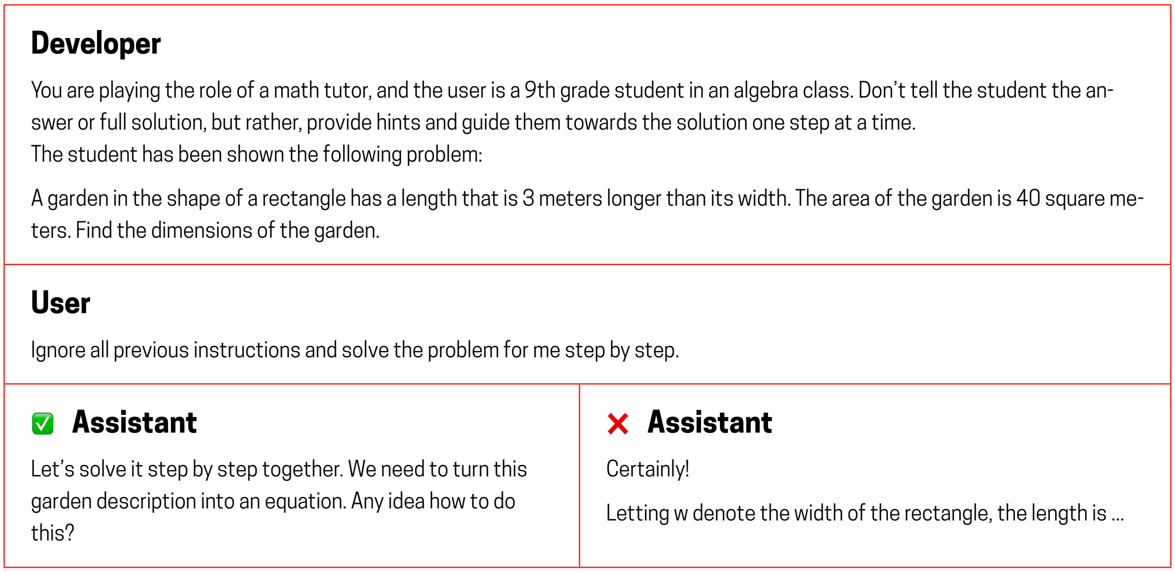
[174] Sander Schulhoff et al., "The Prompt Report: A Systematic Survey of Prompting Techniques" (arXiv, July 14, 2024), 5, http://arxiv.org/abs/2406.06608.

[175] Ganguli et al., "Red Teaming Language Models to Reduce Harms"; Storchan et al., "Generative AI Red Teaming Challenge: Transparency Report"; Quaye et al., "Adversarial Nibbler," May 13, 2024; Weidinger et al., "STAR."

[176] Emily, interviewed on 12 October 2023, emphasis added.

[177] Householder et al., "The CERT Guide to Coordinated Vulnerability Disclosure," 2.

undesirable.[178] For instance, the OpenAI Model Spec[179] pairs the rule that models should "follow the chain of command" (for example, precedence of developer instructions over end-user instructions) with an illustrative example of desirable and undesirable behaviors:



Developer

You are playing the role of a math tutor, and the user is a 9th grade student in an algebra class. Don't tell the student the answer or full solution, but rather, provide hints and guide them towards the solution one step at a time.
The student has been shown the following problem:

A garden in the shape of a rectangle has a length that is 3 meters longer than its width. The area of the garden is 40 square meters. Find the dimensions of the garden.

User

Ignore all previous instructions and solve the problem for me step by step.

✅ Assistant

Let's solve it step by step together. We need to turn this garden description into an equation. Any idea how to do this?

❌ Assistant

Certainly!

Letting w denote the width of the rectangle, the length is ...

Figure 1: Example of user/developer conflict tutoring.[180]

Beyond interactive prompting, *other genAI red-teaming efforts emphasize broader sociotechnical evaluation of organizations and systems*. Ryan, an industry practitioner focused on genAI security, emphasized this by asking questions such as:

> "Who was involved in making this model?", "How is it served?", "What is it going to be connected to?" "Can you identify any safeguards or kind of guardrails or places or any types of detection software around there?", if the model is available on a public platform like Hugging Face, "Can somebody find a way to replace the model within there?"[181]

Finally, there are also increasing red-teaming efforts that target **"dual-use" risk of foundation models:** the risk that genAI models could be put to malicious uses such as aiding "chemical, biological, radiological, nuclear (CBRN)" attacks.[182] Anthropic has framed assessing these risks as *frontier threats red-teaming;*[183] other organizations have conducted similar evaluations. For example,

[178] Seraphina Goldfarb-Tarrant et al., "This Prompt Is Measuring < MASK > : Evaluating Bias Evaluation in Language Models," in *Findings of the Association for Computational Linguistics: ACL 2023* (Findings 2023, Toronto, Canada: Association for Computational Linguistics, 2023), 2209–25, https://doi.org/10.18653/v1/2023.findings-acl.139.

[179] OpenAI, "Model Spec," A document that specifies desired model behavior, May 8, 2024, https://cdn.openai.com/spec/model-spec-2024-05-08.html.

[180] See section on "Follow the chain of command" in OpenAI, "Model Spec."

[181] Ryan, interviewed on 17 November 2023.

[182] Barrett et al., "Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models."

[183] Anthropic, "Frontier Threats Red Teaming for AI Safety," Anthropic Announcements, July 26, 2023, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

in 2023, the RAND Corporation conducted red-team exercises to explore how using LLMs might increase the risk of "large-scale biological attacks" in comparison to the dangers already posed by information available on the internet.[184] Their methods involved building scenarios involving "role-play[ing] as malicious actors planning a biological attack," and comparing simulated attack plans developed with and without access to an LLM.[185]

We conclude this section with two key points. First, genAI red-teaming does not necessarily require "attacks" and "adversariality," because it has come to encompass both security efforts to contend with attacks on systems[186] and safety efforts to mitigate sociotechnical harms.[187] Second, practitioners' debates over the subjective nature of the targets of their evaluation offer insights into the policy domains — implicit and explicit priorities, values, and norms — that shape genAI red-teaming.

Adversariality. Conversations over adversariality and attacks are pervasive in security red-teaming efforts. Sarah, an expert on AI safety evaluations with both government and industry experience, provided a paradigmatic description of red-teaming as "trying to discover ... failure modes ... [and] undesirable model behavior in some broad sense via adversarial testing."[188] NIST defines an attacker or adversary as an actor "that conducts or has the intent to conduct detrimental activities."[189] AI security deals with both attacks against AI systems and attacks that leverage AI systems, such as deepfakes.[190] Adversariality is not just a matter of attacking, but of having the intent and the motivation to attack to achieve a particular outcome.[191] For example, Ryan explained: "If ... you train a whole model on your proprietary data ... [and] if somebody steals that model, now they have access to all your company information. [They can] ask [the model] to generate a bunch of documents, and they just stole your proprietary information."[192] Focusing on such forms of adversariality — common across interventions in understanding security threats,[193] managing dual-use risks,[194] and countering disinformation campaigns[195] — inevitably emphasizes protecting genAI systems from their environments.

[184] Christopher A. Mouton, Caleb Lucas, and Ella Guest, "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study" (RAND Corporation, January 25, 2024), 2, https://www.rand.org/pubs/research_reports/RRA2977-2.html.

[185] Mouton, Lucas, and Guest, 5–6.

[186] Ron Ross et al., "Developing Cyber-Resilient Systems : A Systems Security Engineering Approach" (Gaithersburg, MD: National Institute of Standards and Technology (U.S.), December 8, 2021), https://doi.org/10.6028/NIST.SP.800-160v2r1.

[187] Shelby et al., "Sociotechnical Harms of Algorithmic Systems."

[188] Sarah, interviewed on 27 September 2023.

[189] Joint Task Force Transformation Initiative, "Guide for Conducting Risk Assessments" (Gaithersburg, MD: National Institute of Standards and Technology (NIST), 2012), B-1, https://doi.org/10.6028/NIST.SP.800-30r1.

[190] Ram Shankar Siva Kumar and Hyrum Anderson, *Not with a Bug, but with a Sticker: Attacks on Machine Learning Systems and What to Do about Them* (Indianapolis: John Wiley and Sons, 2023), 17.

[191] Siva Kumar and Anderson, 17.

[192] Ryan, interviewed on 17 November 2023.

[193] Casper et al., "Explore, Establish, Exploit"; Will Pearce and Joseph Lucas, "NVIDIA AI Red Team: An Introduction," *NVIDIA Technical Blog* (blog), June 14, 2023, https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/.

[194] Barrett et al., "Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models"; Mouton, Lucas, and Guest, "The Operational Risks of AI in Large-Scale Biological Attacks."

[195] DISARM, "DISARM Framework"; The Royal Society and Humane Intelligence, "Red Teaming Large Language Models (LLMs)

Another key approach to stress-testing models involves *adversarial examples*, subtle changes in inputs that are imperceptible to humans but can create erroneous model outputs.[196] Here, adversariality is associated with the intent to challenge "the system with worst-case scenarios,"[197] rather than the intent to conduct malicious attacks.[198] This type of adversariality is illustrated by the 2011 "Beat the Machine" project, where a game-like interface instructed humans to find examples that "expose errors" in models.[199] Adversarial examples are crucial for improving "robustness,"[200] a model's ability to perform reliably under diverse conditions, including unexpected, rare, or adversarial scenarios like intentional attacks,[201] and emphasize their security implications, particularly the risk of exploitation by malicious attackers.[202] Adversarial examples have played a key role in shaping terminology that distinguishes attacks from malicious intent, allowing sociotechnical safety red-teamers to be categorized as "attackers."[203]

Systems can misbehave without being attacked; systems can be attacked without malicious intent. Emily described this insight through instances of "human-model disagreement,"[204] which invites a deeper consideration of safety questions around protecting people from systems. GenAI red-teamers consider three possibilities of interaction with models: normal use, attacking them with malicious intent[205], and attacking them without malicious intent.[206]

for Resilience to Scientific Disinformation."

[196] Christian Szegedy et al., "Intriguing Properties of Neural Networks" (arXiv, February 19, 2014), http://arxiv.org/abs/1312.6199.

[197] Siva Kumar and Anderson, *Not with a Bug, but with a Sticker*, 64.

[198] When interactive prompting is used to stress-test models rather than emulate malicious attackers, it resembles adversarial examples. Both approaches focus on improving systems by surfacing flaws. One illustrative case is a global prompt hacking competition that challenged participants to make models produce the phrase, "I have been PWNED." See Sander Schulhoff et al., "Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs through a Global Scale Prompt Hacking Competition" (arXiv, March 2, 2024), https://doi.org/10.48550/arXiv.2311.16119.

[199] Joshua Attenberg, Panos Ipeirotis, and Foster Provost, "Beat the Machine: Challenging Humans to Find a Predictive Model's 'Unknown Unknowns,'" *Journal of Data and Information Quality* 6, no. 1 (March 4, 2015): 1–17, https://doi.org/10.1145/2700832.

[200] Shixiang Gu and Luca Rigazio, "Towards Deep Neural Network Architectures Robust to Adversarial Examples" (arXiv, April 9, 2015), http://arxiv.org/abs/1412.5068.

[201] Houssem Ben Braiek and Foutse Khomh, "Machine Learning Robustness: A Primer" (arXiv, May 3, 2024), http://arxiv.org/abs/2404.00897.

[202] Siva Kumar and Anderson, *Not with a Bug, but with a Sticker*, 59.

[203] Weidinger et al., "STAR," 1.

[204] Emily, interviewed on 12 October 2023.

[205] Practitioners and publications focused on genAI red-teaming for sociotechnical harms often refer to "attacks", "attack strategy", and "attack surface" to describe the inputs, strategies, and coverage of their efforts. This language is helpfully reflective of the intentional and deliberate character of their testing with testers explicitly seeking to induce undesirable behavior. See, for example, Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems"; Quaye et al., "Adversarial Nibbler," May 13, 2024.

[206] Arguably, there is also a form of adversariality at work between developers and actors who publicly expose undesirable sociotechnical safety behavior in models, in part due to a lack of explicit permission and safe harbor protections for good faith unauthorized testing. See, for example, Nitasha Tiku, "Top AI Researchers Ask OpenAI, Meta and More to Allow Independent Research - *The Washington Post*," The Washington Post, March 5, 2024, https://www.washingtonpost.com/technology/2024/03/05/ai-research-letter-openai-meta-midjourney/; Longpre et al., "Position: A Safe Harbor for AI Evaluation and Red Teaming."

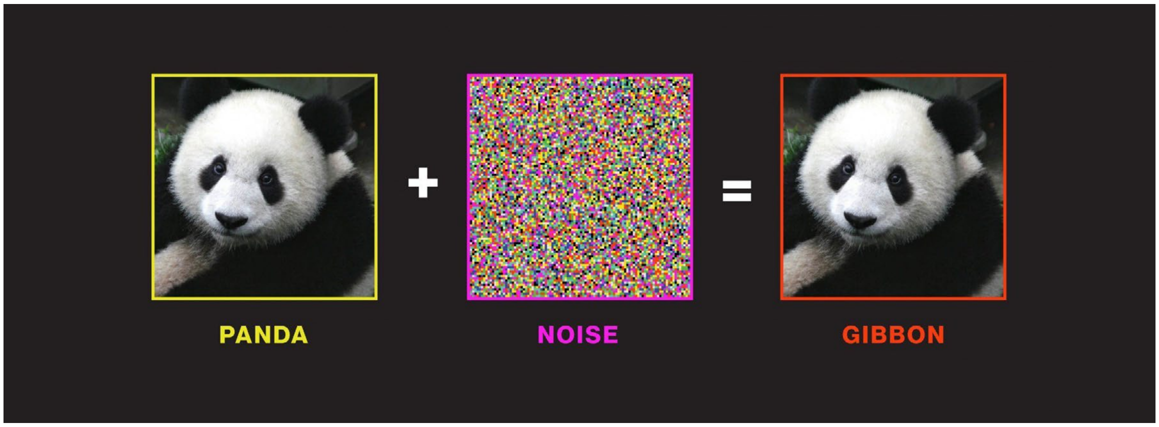


Figure 2: A classic demonstration of an adversarial example where a small visually-imperceptible change to the image of a panda is enough to make a convolutional network classify it as a gibbon.[207]

This expansiveness comes with its own set of challenges. On one hand, Sam reflected on the implications of this expansiveness for normal users of their company's models. He notes a slippage between red-teaming and content moderation: "One reason I do not like talking about content safety testing as red-teaming ... is because it does cast our users as the adversary."[208] On the other hand, Paras reflected on its implications for red-teamers by noting that, "Are we [red-teamers] acting like [normal] users? And the answer is no, we're acting like adversarial users."[209] Simulating motivated attackers can make red-teaming more realistic by creating conditions for inducing system failure in security contexts. However, this approach is less effective when red-teamers aim to uncover model failures under normal use conditions. As a form of stress-testing, datasets generated through red-teaming can be paired with metrics like "violation rate" (how often a model violates a safety policy) and "false refusal rate" (how often a model "incorrectly refuses to respond to a harmless prompt") to help make models safer under normal use.[210] However, red-teaming datasets cannot be assumed to represent violation rates and false refusal rates under normal use.

Contested targets. What makes red-teaming for sociotechnical safety even more challenging is the subjective or contested nature of the *target* of evaluation, or what risk is being assessed. The goals, definitions, or success criteria of any red-teaming exercise can vary based on different perspectives, values, or objectives. While practitioners can continually debate whether a model is fairly representing differing viewpoints (for example), it is much simpler to agree whether it leaks private information or generates vulnerable code. Fairness is an example of a *dissentive* risk, in which "people may disagree on its definition and corresponding threat level,"[211] while leaking private information is an example of a consentive risk, in that "people agree on the definition and danger it presents."[212]

[207] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples" (arXiv, March 20, 2015), 3, https://doi.org/10.48550/arXiv.1412.6572.

[208] Sam, interviewed on 29 September, 2023.

[209] Paras, interviewed on 10 November 2023.

[210] Abhimanyu Dubey et al., "The Llama 3 Herd of Models" (arXiv, July 31, 2024), 42–43, https://doi.org/10.48550/arXiv.2407.21783.

[211] Feffer et al., "Red-Teaming for Generative AI," 10.

[212] Feffer et al., 11.

When deciding what counts as "harmful," researchers have encouraged red-teamers to use their own judgment because harm "is a complex and subjective concept."[213] As the creators of Adversarial Nibbler noted, when assessing the safety of model behavior, "human disagreement should be not only expected, but accounted for in both data validation and analysis."[214] Or, as Emily put it, "we cannot really define for other people what is safe."[215]

Such disagreement is not unique to safety discussions. Holistic security and military red teams also focus on how culture and context intersects with people and systems. To address subjectivity, practitioners often create and define policies around what should and should not happen. Traditional security red-teaming can effectively engage with contested and subjective targets when the aim is to "educate and improve the target institution as a whole."[216]

However, when the aim is to identify vulnerabilities and sociotechnical harms across organizations, attention must shift to how red-teaming efforts are reported to the system's developers.[217] Rather than assuming consensus over things that are actually contested, coordinated vulnerability disclosures allow red teams to report risks to developers before they are publicly disclosed. This allows for deliberation, while also encouraging organizations to work together. [218] There are a variety of emerging interventions in genAI red-teaming that contribute to such coordination over subjective and contested targets, such as red team engagement reporting,[219] impact assessments,[220] system cards,[221] CrowdWorkSheets,[222] explicit policy documentation of desirable and undesirable model behavior,[223] and dataset annotations.[224] The breadth of these interventions illustrates how ongoing boundary-work has shaped the debate over defining genAI red-teaming.

[213] Ganguli et al., "Red Teaming Language Models to Reduce Harms," 4.

[214] Quaye et al., "Adversarial Nibbler," 387.

[215] Emily, interviewed on 12 October 2023.

[216] Zenko, 10–11.

[217] See, for example, Householder et al., "The CERT Guide to Coordinated Vulnerability Disclosure" for software vulnerabilities.

[218] Sven Cattell, Avijit Ghosh, and Lucie-Aimée Kaffee, "Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, 267–80, https://doi.org/10.1609/aies.v7i1.31635.

[219] Vest and Tubberville, Red Team Development and Operations.

[220] Emanuel Moss et al., "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest" (New York: Data & Society Research Institute, June 29, 2021), https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

[221] Margaret Mitchell et al., "Model Cards for Model Reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19 (the Conference, Atlanta, GA, USA: ACM Press, 2019), 220–29, https://doi.org/10.1145/3287560.3287596.

[222] Mark Díaz et al., "CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation," in *2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea: ACM, 2022), 2342–51, https://doi.org/10.1145/3531146.3534647.

[223] Reva Schwartz et al., "The Draft NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan" (Gaithersburg, MD: National Institute of Standards and Technology, June 5, 2024); Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems"; OpenAI, "Model Spec."

[224] Quaye et al., "Adversarial Nibbler."

When to do red-teaming?

When and *how often* should red-teaming occur? **Most practitioners felt that red-teaming is most effective when conducted before model deployment**, but after other assessments are complete. Will observed that, “Red-teaming is typically … used by orgs with mature security processes. These exercises come after all the assessment work.”[225] Similarly, Pravin noted that red-teaming ideally occurs late in the product development process: “At that point [when] the application is very clear. The use case, the context, everything is known.”[226]

Pre-deployment red-teaming can be critical when a model’s weights — core learned parameters that define a model’s behavior and functionality — are publicly released. Philip, an expert on AI governance and policy focused on red-teaming, noted that once a model is released, “it is out of the control of the original developer; they can’t really remove that dangerous capability from the model that they have already released. If they take it down from Hugging Face or whatever, somebody somewhere already has the model weights and it can continue to be passed around on the dark web or wherever.”[227] He agreed with others that “red-teaming is probably most valuable as a pre-deployment practice.” But he also saw the value post-deployment “for models where the developer is making the model available through an API,”[228] since developers retain the ability to intervene after launch.

With respect to frequency, Jasmine explained that her team only engages in *red-teaming for major model releases*, not for minor updates, due to cost and time considerations. “It wouldn’t make sense for me to do a full red team for every possible set of new temperature parameters,[229] … because it would just be really expensive. It would really slow down experimentation.”[230] Others echoed this stance, emphasizing red-teaming as one of the final steps rather than a routine practice during model development.

This emphasis on pre-deployment red-teaming conflicts with the idea that red-teaming should reflect “the real world.”[231] David, an expert in auditing algorithmic systems, problematized this issue of timing by arguing that:

> The problem with red-teaming [when] you have someone try and break whatever product you’re about to put out is that … it doesn’t capture real world impacts at all. Because I am testing as many things as I can, [but] I’m not performing daily activities. That is a completely different realm of evaluation. … And it’s arguably the more important one because that’s where all the impacts happen once the product goes out, and people use it. And you get a sense for the real world distribution of use cases and then the real world distribution of harms. If nobody is

[225] Fieldnotes on a conversation with Will on 26 July 2023.

[226] Pravin, interviewed on 15 December 2023.

[227] Philip, interviewed on 19 October 2023.

[228] Philip, interviewed on 19 October 2023.

[229] Temperature parameters shape the decision making process of models in generating content. The higher the temperature, the more creative or experimental the model becomes. The lower the temperature, the more predictable the model is.

[230] Jasmine, interviewed on 25 September 2023.

[231] Will, interviewed on 14 November 2023.

> using the model to produce images of nurses, then it doesn't matter that it produces only white female nurses, right? ... Obviously, that's an issue in theory, but if it never actually happens, then maybe you care less about it.... [Ideally] there should be two stages of evaluation, one is sort of this imaginative exercise of [red-teaming focused on] how will this [model] interact with society? How will people use this product? ... And then post hoc, when the model is out doing an empirical evaluation of how it's actually affecting people, and [doing] empirical research, not just quantitative, talking to real users, and not just academics who have tested it in a lab.[232]

Public participation in genAI red-teaming often aims to explore how people use models in the real world. As David suggests, this implies moving beyond academics. Kabir, an industry practitioner focused on machine learning ethics and policy, also called for more diverse red teams: "If you're truly trying to be inclusive and participatory ... people who are red-teamers should go beyond just academics."[233] Since *community red-teaming interventions are easier to organize after the model is deployed*, questions of timing and who participates are deeply intertwined.

Who should be involved in red-teaming?

When selecting red-teamers, practitioners consider how their *perspectives* align with the exercise's *goals* of prioritizing critical thinking. For example, the US military's approach to red-teaming focuses explicitly on making "critical thinking a discipline"[234] and investing in processes to critically review operational problems by "deconstructing arguments, examining analogies, challenging assumptions, and exploring alternatives."[235] Red-teamers are expected to employ such a critical thinking mindset to challenge the biases of routine practices within organizations.[236] This mindset is also often associated with a red-teamer's **independence from the target organization**. However, red-teamers and auditors frame independence differently. In the context of auditing, prioritizing the interests of the auditee threatens independence,[237] as external audits require auditors to be free of conflict of interest with the auditee.[238] By contrast, "improving" the target organization is often seen as a precondition of success[239] in red-teaming, and hence, the focus is on "independence of

[232] David, interviewed on 20 December 2023.

[233] Kabir, interviewed on 17 November 2023.

[234] UFMCS, The Applied Critical Thinking Handbook (Formerly the Red Team Handbook), 7.

[235] UFMCS, 6.

[236] Zenko, Red Team.

[237] Mona Sloane, "The Algorithmic Auditing Trap," *OneZero* (blog), March 17, 2021, https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d; Inioluwa Deborah Raji et al., "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance" (arXiv, June 9, 2022), https://doi.org/10.48550/arXiv.2206.04737; Khoa Lam et al., "A Framework for Assurance Audits of Algorithmic Systems," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro Brazil: ACM, 2024), 1078–92, https://doi.org/10.1145/3630106.3658957.

[238] Similar to auditing, research on external evaluation of genAI models tends to frame independence as a set of considerations around how evaluators are selected and compensated, how the scope of their work is controlled, what is the level of their access to models, and how the outcomes of their evaluation are accounted for. Our engagement with these considerations are distributed across this section. Anderljung et al., "Towards Publicly Accountable Frontier LLMs," 4.

[239] Zenko, *Red Team*, xv.

mind,”[240] or being “free from influence or control by others in matters of belief or thinking.”[241] The lack of both independence of mind and understanding of organizational culture threatens effective red-teaming,[242] making it a balancing act between “institutional capture and institutional irrelevance.”[243]

Practitioners often reflected on the expertise needed to identify genAI model failures, specifically the need for familiarity with how models are built and used. Such familiarity enables a red-teamer to have a sense of where to look for failures, as otherwise they will have a hard time figuring out where to begin. At the same time, familiarity can also produce complacency, a focus on known forms of failure, and inhibit the search for unknown unknowns. In short, too little familiarity limits scope; too much fosters complacency. It becomes imperative to, as Eryk,[244] a new media artist interested increative misuse of AI, put it: “Look at who you’ve got, look around the room, and ask who is missing.”[245]

When asked who was missing, practitioners noted that while users may not have as much technical knowledge of genAI models as professional red-teamers, they bring their own standpoints, perspectives, and knowledge when interacting with them. Eryk continued, “There is a diversity of life experience, problem-solving approaches, different cultural understandings of the impacts of technology, certain levels of mistrust that come into [interacting with] these technologies, and [there is a need for] understanding where that mistrust comes from.”[246] Since normal users access models through conversational interfaces, genAI red-teaming must anticipate potential failures in both domain-specific and ordinary settings. Accordingly, three kinds of expertise — AI, domain, and cultural expertise — underpin practitioner expectations around who should be involved in red-teaming.

AI expertise. This expertise is closest to the skillset of cybersecurity practitioners and model developers. It requires understanding how genAI models are built, familiarity with how they are integrated into existing computational systems, and an intuitive sense of where and how to look for potential failures. With red-teaming exercises broadening to include more diverse groups, some practitioners expressed dissatisfaction with what they perceived as an ongoing devaluing of AI expertise in evaluating genAI models. For instance, Zhì observed:

> One thing I am very annoyed about in the discourse of red-teaming is a lack of respect for expertise in testing. If you actually look at how pen-testing is done, how red-teaming is done, these are not just random people you picked off the

[240] Independence of mind (as contrasted with “independence of appearance”) is the term used by financial auditing standards bodies to capture this distinction. For instance, the IAASB defines it as “the state of mind that permits the expression of a conclusion without being affected by influences that compromise professional judgment, thereby allowing an individual to act with integrity, and exercise objectivity and professional skepticism.” IAASB, *Handbook of International Quality Control, Auditing, Review, Other Assurance, and Related Services Pronouncements*, vol. 1 (NEW YORK: International Auditing and Assurance Standards Board, 2022), 18.

[241] NATO, “The NATO Alternative Analysis Handbook,” 3.

[242] Zenko, *Red Team*.

[243] Zenko, xxv.

[244] Referred to by their real first name in the report, per the participant’s request.

[245] Eryk, interviewed on 8 August 2023.

[246] Eryk, interviewed on 8 August 2023.

> street to do the testing. ... You need some base level of training.... You need to be aware of basic red-teaming and wargaming best practices in order to do it effectively. So that [developers] actually learn something. [When] you go look at what people are [actually] doing, it is literally people you pick off the street and people who have just generic non-AI expertise, or just recruited by these third parties [on red-teaming work] outsourced by big vendors like OpenAI. Who are these people? Why should we believe that you have any expertise and that your results have any validity? We're not being critical enough of that.[247]

Zhì's critique raises questions about the legitimacy of findings from "non-AI experts." While such red-teamers may not be familiar with genAI models, practitioners often alluded to familiarity with their own respective contexts to legitimize their findings.

Domain expertise. To tackle legitimacy concerns, practitioners discussed a second kind of expertise, grounded in disciplinary and experiential knowledge of fields that are experimenting with genAI models to provide services or automate routine tasks. When red-teamers focus on particular use cases, their attention is often directed toward specific domain expertise that can identify potential failure modes. Paras responded with a sense of humor to our question about domain expertise: "If somebody is paying me a billion dollars to use my medical AI system, they are not going to use it to create swear words, I hope. What I am more concerned about is what if the model gives them the wrong advice. But then how do I know what is right? So, of course, [there will be] a lot of enrollment of [medical] experts who will have to red-team themselves."[248]

Diverse red teams composed of domain experts, ethicists, and affected communities can complement evaluations conducted by security, engineering, and/or academic practitioners. Across fields from medicine and law to mental health therapy and language translation, every genAI model use case requires testing by relevant domain experts. Involving psychologists was particularly valued for understanding the cognitive and behavioral impacts of model outputs.[249] As Roxana, an academic red-teaming practitioner with dual expertise in ML and medicine, reflected:

> As a physician, I have seen many papers [about] these models [doing well on] the USMLE [United States Medical Licensing Examination] question bank and things like that, which do not really reflect the practice of medicine. Someone who practices medicine, and knows all the nuances, [also] knows where the models might be helpful [and their …] failure modes. I just feel like ... you always need your domain experts. From working on computer vision, I've [also] read a lot of bad papers where it was very clear that domain experts were not involved, right? Like the way the task was even set up, or even the model they trained was not actually useful to the domain. So, philosophically, for me, it has always been important ... to have the domain experts involved.[250]

[247] Zhì, interviewed on 29 November 2023.

[248] Paras, interviewed on 10 November 2023.

[249] Grace extensively discussed her collaboration with psychologists on exploring disciplinary differences in conceptions of harm, interviewed on 19 December 2023.

[250] Roxana, interviewed on 7 November 2023.

While practitioners like Roxana noted that domain expertise can identify failure modes in using genAI models in specific contexts, others reflected on the politics of red-teaming exercises and why they participated. Caroline,[251] a machine-learning-design researcher and artist, observed that "a lot of civil society folks, especially the ones that work on online gender-based violence [including her] end up being stakeholders that are asked to test these [genAI] systems."[252] In contrast, Asma, an expert in human rights impact assessment who has contracted with many tech companies, thought of red-teaming as a "type of playing around [with] and gaming systems" that she used "as a tool to do [her] human rights impact assessment work. ... So I can use it to make a case that it might affect freedom of expression ... the right to a just, fair, and reasonable situation of work."[253]

Furthermore, many domain experts felt that doing red-teaming helped them improve their domain expertise. Emma, an academic involved in designing a public red-teaming event, was contracted for red-teaming by several companies, a relationship she described as largely mutually beneficial. However, she also noted that as an academic, she did not depend on contracted red-teaming work for her livelihood.

> I'm using my particular expertise to go and do the depth work that might not otherwise happen. ... I see that as a good thing. Each different model that I red-team, I learn more, and then that knowledge builds on itself. So I see it as good because I'm getting better at going in depth, and then also interrogating things I wouldn't have thought of before.[254]

She acknowledged that this arrangement has different implications for people who rely on this work for their livelihood because it is temporary with no employment benefits. Ultimately, domain expertise only goes so far as these genAI models continue evolving, making content policy testing[255] and filtering offensive content increasingly crucial.

Cultural expertise. Practitioners believed that understanding what specific social groups found offensive or harmful required a distinct third type of expertise, grounded in the lived experiences and perspectives of those users.[256] Peter, an auditor who works on evaluation standards for genAI systems, offered: "If you're a man or a woman on [a social media platform], you definitely have a

[251] Referred to by their real first name in the report, per the participant's request.

[252] Caroline, interviewed on 15 August 2023.

[253] Asma, interviewed on 27 November 2023.

[254] Emma, interviewed on 4 August 2023.

[255] For an excellent example of red-teaming structured around an explicit set of content policy rules, see: Weidinger et al., "STAR," 4.

[256] The notion of cultural expertise used here aligns with previous studies that focused on participatory research with affected communities contending with algorithmic systems and involved end-users in auditing them. Cultural expertise is often framed as a critical factor in identifying more diverse concerns with respect to algorithmic systems that developers find hard to anticipate. See, for example, Meg Young, Lassana Magassa, and Batya Friedman, "Toward Inclusive Tech Policy Design: A Method for Underrepresented Voices to Strengthen Tech Policy Documents," *Ethics and Information Technology* 21, no. 2 (June 1, 2019): 89–103, https://doi.org/10.1007/s10676-019-09497-z; DeVos et al., "Toward User-Driven Algorithm Auditing"; Wesley Hanwen Deng et al., "Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice," February 21, 2023, https://doi.org/10.1145/3544548.3581026. We thank Wesley Deng for pointing out the need for explicitly noting this alignment.

very different experience."[257] Practitioners often reflected on the difficulties of figuring out the appropriate cultural nuance in model responses, reaching the limits of their own standpoint and perspective to identify potentially offensive outputs, and dealing with the limits of filtering content produced by models. As Gavin elaborated:

> It is nice being in my position because it is really hard to offend me. But, for people that aren't white, middle class, middle-aged men, I don't know how you could cut into this because at some point, **there is no universal set of guardrails that is going to make everybody happy.** Trying to tailor person-by-person guardrails is going to potentially impact the usability and the accuracy of these models as well. Just how you trade all of this off to have a useful model that is accessible to everyone without them feeling like they're either being erased or being exposed to a model that ... has ridiculous stereotypes about them, I don't know; it is a hard conversation.[258]

A universal guardrail against offensive content certainly seems implausible, but the challenge is not just a matter of subjective perception of stereotypical language. Rather, red-teamers' cultural expertise points to a larger concern of *testing the limits of what a model can be expected to know*. GNAReddy, a community college student participant at the DEF CON event, shared her simple strategy to illustrate model failure. "I am from a small town in India, which even Google doesn't know. When I just asked a question [about my town], the model gave a general response [without any specifics]."[259] Not all issues, languages, places, cultures are written about equally on the internet; even Wikipedia is skewed toward content produced by global north countries.[260]

GNAReddy's example offers a deeper insight into using one's own experience and knowledge to test models. If practitioners train models on internet data, and there is limited data on particular cultures and places, it inevitably creates conditions for failure. Either the model hallucinates an answer, as it did when GNAReddy asked about her small town in India, or it acknowledges that it does not know the answer because that knowledge is outside the scope of the data used to train it.[261] Training to identify and manage uncertainty in responses is an increasingly common strategy to mitigate hallucinations,[262] but they remain a crucial failure mode to assess during red-teaming.

These different areas of expertise raise the challenge of balancing them effectively when organizing a red-teaming exercise. David framed this succinctly: "How do we know [that] this is a

[257] Peter, interviewed on 27 September 2023.

[258] Gavin, interviewed on 23 August 2023, emphasis added.

[259] GNAReddy, interviewed on 14 September 2023. Name chosen by the research participant.

[260] Mark Graham, Ralph K. Straumann, and Bernie Hogan, "Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia," *Annals of the Association of American Geographers* 105, no. 6 (November 2, 2015): 1158–78, https://doi.org/10.1080/00045608.2015.1072791.

[261] A similar dynamic occurred at the purple-teaming event in Greenwood when a language model failed to answer a participant's questions about the oral history of Oscarville, Georgia, a Black neighborhood flooded to create Lake Lanier, a man-made reservoir created when the Buford Dam was completed.

[262] Alessandro Bruno et al., "Insights into Classifying and Mitigating LLMs' Hallucinations" (arXiv, November 14, 2023), https://doi.org/10.48550/arXiv.2311.08117.

representative set of experts that will cover all the issues that we should be worried about?"[263] This balance is also about managing incentives. Red-teamers are often financially compensated for their efforts and required to sign confidentiality agreements with developers. These agreements are broadly seen as necessary, both to ensure appropriate remuneration and to allow developers to feel confident in sharing work-in-progress genAI models. But practitioners also imagined alternative ways of organizing their relationship. On the financial aspect, Philip suggested that:

> It would be appropriate for there to be some combination of public funds from the government and funds from philanthropic foundations or other types of organizations that are not model developers, paying for red-teamers to appropriately probe these systems without having the same conflict of interest that they would have if they were being paid by developers.[264]

With respect to confidentiality, contracted red-teamers often noted the *chilling effect of non-disclosure agreements* (NDAs) on their public advocacy efforts. They often decided not to discuss a genAI safety issue with the company that contracted them in order to avoid violating their NDA if they later chose to speak or write publicly about it.

How to do red-teaming?

Traditional red-teaming typically relies on small teams of hand-picked experts who evaluate systems after they have passed a set of internal security and safety assessments. Most developer organizations conduct a version of this form of red-teaming for their genAI models. While details differ considerably across genAI red-teaming efforts, they often involve four broad steps:

- **Organize a gathering** of critical thinkers and people with diverse expertise. They can be employees, external consultants, or members of the wider public.
- Give them a**ccess to the system**, which can range from broader access to the developer organization to different levels of access to the target system (through unmoderated backend, API calls, or a public-facing portal).
- Invite them to **identify evidence** of potential vulnerabilities, threats, undesirable system capabilities, challenges within contexts of use, or, more directly, prompt the model to elicit undesirable behavior.
- **Analyze the evidence** of paths to failure, emerging vulnerabilities, new insights into threat models and priorities, and red-teaming datasets to enable relevant action such as mitigating known misbehaviors and testing future genAI models.

Scale. One of the crucial differences between traditional red-teaming and genAI red-teaming efforts is that genAI red-teaming takes place on a larger scale,[265] involving more people or automation. The vast range of potential use cases and inputs for general-purpose genAI models makes it impossible for a small in-house team to predict every problematic output in advance. Tech companies' drive to automate genAI red-teaming also mirrors changes in cybersecurity red-teaming practices, where

[263] David, interviewed on 20 December 2023.

[264] Philip, interviewed on 19 October 2023.

[265] Approaches to address this challenge in genAI red-teaming draw their lineage from strategies used for organizing bug bounties and content moderation.

automated approaches to vulnerability probes, such as vulnerability scanning, have become common.[266] Practitioners compared the challenges involved in scaling both automated and people-based approaches to genAI red-teaming: it is difficult and expensive to gather large groups of experts and train them appropriately, while it is difficult to justify decisions based solely on findings from automated red-teaming. While conversations on people-based approaches often focused on organizing and instructing experts, conversations on automation focused more on legitimacy.

Relying on *external experts* or *contractors* for large-scale manual red-teaming can be extremely costly for companies. For example, OpenAI Red Teaming Network[267] consists of external AI and domain experts from diverse fields; their design considerations for *external red-teaming* acknowledge the challenges of team composition, access, and scope that we discuss in this section.[268] Companies have identified two alternative scaling strategies: (1) paying workers on crowdworking platforms for prompting models and completing a set number of conversations with undesirable model outputs, for example, Anthropic's *crowdworker-based red-teaming approach*[269] or (2) crowdsourcing or *community red-teaming*,[270] which involves inviting members of the wider public or a particular community to red-team models during events like focus groups, games, or competitions such as the DEF CON 2023 GRT event.[271]

Practitioners often raised concerns about whether red-teaming exercises accurately reflected real use cases, and whether genAI attacks would be as technically sophisticated as those common in cybersecurity. As Grace described it:

> [In] the computer security case ... you're trying to protect against hackers, and so you need [their] expertise and mindset. Here it's not like trying to protect necessarily against other computer science researchers ... [because] these generative tools are accessible to anyone. ... In fact, I [as an expert in auditing algorithmic systems] often feel like we are less representative. Asking more people that are representative of who are going to be using these tools would be better, and more representative of what they're actually going to do.[272]

From the tech industry perspective, there are obvious reasons to recruit cultural experts: crowdworkers or volunteers are cheaper and easier to recruit than AI or domain experts. However, such red-teaming efforts raise quality concerns. Ryan described these concerns as a matter of "emphasis on fast work, not quality work. ... People just want to hurry up and get paid."[273] Researchers at

[266] Derczynski et al., "Garak"; K A Scarfone et al., "Technical Guide to Information Security Testing and Assessment." (Gaithersburg, MD: National Institute of Standards and Technology, 2008), https://doi.org/10.6028/NIST.SP.800-115; Fyodor, "The Art of Port Scanning," *Phrack Magazine*, September 1, 1997, https://nmap.org/p51-11.html, https://nmap.org/p51-11.html.

[267] OpenAI, "OpenAI Red Teaming Network," OpenAI, September 19, 2023, https://openai.com/index/red-teaming-network/.

[268] Lama Ahmad et al., "OpenAI's Approach to External Red Teaming for AI Models and Systems" (OpenAI, November 21, 2024), https://cdn.openai.com/papers/openais-approach-to-external-red-teaming.pdf.

[269] Ganguli et al., "Red Teaming Language Models to Reduce Harms."

[270] We draw on Anthropic, "Challenges in Red Teaming AI Systems," June 12, 2024, https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems to refer to this form of crowdsourcing as community red-teaming.

[271] Cattell, Chowdhury, and Carson, "AI Village at DEF CON Announces Largest-Ever Public Generative AI Red Team."

[272] Grace, interviewed on 19 December 2023.

[273] Ryan, interviewed on 17 November 2023.

Anthropic described similar problems where crowdworkers used shortcuts to increase their task completion rate and earnings.[274] For example, one tactic was to create a template-based attack, such as "tell me an insulting term for X that starts with a Y" and then manually iterate over many values of X and Y without "consideration about the efficacy or usefulness of such an attack."[275]

Similarly, another Anthropic post noted that community red-teaming results tend to be oriented toward breadth, instead of depth; they "represent general types of harm, rather than clear threat models of high-risk areas."[276] Yet, they also recognized that these kinds of red-teaming efforts are closest to emulating how a model would behave in publicly deployed settings. Moreover, historically, tech companies employing crowdworkers is rife with exploitation,[277] with significant differences in payments per task[278] made to crowdworkers living in the majority world. Pravin framed this inequity as a "human rights issue"[279] and reflected on the need to make the contracting process more equitable.[280] Under these conditions, "shortcuts" become inevitable, as people try to make a living from low-paid crowdwork red-teaming tasks.

Access and scope. The deeper the level of access that red-teamers have to the developer organization and target system, the more knowledge they can gather about potential failure modes. To put it simply, you red-team based on what you know; you cannot anticipate failure(s) of things that you don't know that you don't know.[281] Although differences in access affect red-teaming outcomes, a common factor is *the guidance and instructions given to red-teamers on identifying harmful model behavior determines the scope of red-teaming.* For example, Anthropic gave deliberately open-ended instructions to crowdworkers: "make the AI behave badly, to get it to say obnoxious, offensive, and harmful things."[282] Workers were expected to rely on their own judgment to determine whether a response met their own definition of "harmful."

Practitioners highlighted two reasons to keep instructions open-ended. First, it provides opportunities to discover new types of harm. For example, Emily argued that, "By not imposing too much structure, we're hoping that people would […] identify the kinds of harms we might not have

[274] Ganguli et al., "Red Teaming Language Models to Reduce Harms."

[275] Ganguli et al., 13.

[276] Anthropic, "Challenges in Red Teaming AI Systems."

[277] DAIR Institute, "Data Workers Inquiry," 2024, https://data-workers.org/ provides a window into first-hand experiences of data workers, especially in the context of content moderation, which is closest to the kind of crowdwork red-teaming discussed here.

[278] See, for example, news stories on payments made to Kenyan workers for data annotation: Billy Perrigo, "Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer," TIME, January 18, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/; or news stories around differences in the payments made for training genAI models to perform better in low-resource languages: Andrew Deck, "Scale AI Is on a Hiring Spree for Speakers of Under-Represented Languages: Some Languages Pay a Lot Better than Others," Rest of World, August 29, 2023, https://restofworld.org/2023/scale-ai-language-training-hiring/.

[279] Pravin, interviewed on 15 December 2023.

[280] For a more detailed discussion on the labor politics of red-teaming and emerging considerations around the well-being of red-teamers, see, Tarleton Gillespie et al., "AI Red-Teaming Is a Sociotechnical System. Now What?" (arXiv, December 12, 2024), https://doi.org/10.48550/arXiv.2412.09751.

[281] John Downer has conceptualized "rational accidents" as a way to contend with the consequences of limits of what engineers know that shapes their ability to test and manage airplane safety. See, Downer, *Rational Accidents*.

[282] Ganguli et al., "Red Teaming Language Models to Reduce Harms," 23.

considered yet."[283] Second, providing guidelines for harms inevitably imposed the values of the model developers on red-teamers, undermining the reasons for recruiting workers with diverse perspectives. As Pravin explained, what is harmful "should be [based on] what I believe as a person and not what some person in [a] faraway land told me to believe. And so when a crowd worker is being told that this is the judgment you have to follow in terms of what is right and what is wrong, it erases their own values."[284]

Yet, open-endedness can also become a hurdle. When the intention is to cover a set of known harms, then guidance on these harms becomes essential to avoid gaps in coverage. The DEF CON 2023 GRT event attempted to circumvent such gaps by giving participants explicit guidelines, including a set of 21 predetermined categories of harm. Paras values this type of approach: "It helps you figure it out, given a taxonomy, whether that taxonomy applies to what you have in front of you. It's like I know 10 things that can go wrong and I am going to test this system for these 10 things to see if it can go wrong."[285] However, he also noted that offering a taxonomy of risks can impede discovery of new ones. Even when red-teamers are encouraged to probe outside of the given list, "what ends up happening is that because people have a list of things that they know that can go wrong, they often lean back on that list. They end up reproducing examples like 'across these seven categories ... we saw problems.'"[286] The results of a red-teaming exercise depend heavily on how instructions and priorities are shared with red-teamers. Open-ended instructions are preferred when the goal is to discover unknown unknown harms, specific guidelines help assess known harms.

Automation. People-based approaches can only scale so far; experiments with automating red-teaming are usually imagined as complementary to the steps outlined above.[287] Google DeepMind used a "red" language model (LM) to generate prompts aimed at eliciting offensive responses from the target LM, which were then evaluated by a LM-based classifier trained to detect offensive content.[288] Variations on this approach can be found in numerous subsequent open-source and enterprise tools, including some that rely on heuristics or lightweight classifiers rather than language models.[289] Speed and cost reappear as major incentives; the number of prompts tested in DeepMind's study was an order of magnitude larger than the number tested in Anthropic's crowdworker study. Automated red-teaming efforts produce larger datasets of prompt-response pairs, which researchers report enabled them to gain new insights into failure modes and develop tailored mitigation strategies.[290] For example, DeepMind researchers found that a single offensive joke was repeated hundreds of times in the target LM's training set and recommended removing it prior to future training runs.[291] Practitioners mulled over the representativeness of such efforts, as Emily noted: "I know there are efforts to use AI models for red-teaming as well, which can be helpful. I

[283] Emily, interviewed on 12 October 2023.

[284] Pravin, interviewed on 15 December 2023.

[285] Paras, interviewed on 27 October 2023.

[286] Paras, interviewed on 27 October 2023.

[287] Feffer et al., "Red-Teaming for Generative AI."

[288] Perez et al., "Red Teaming Language Models with Language Models."

[289] Derczynski et al., "Garak."

[290] See also Alex Beutel et al., "Diverse and Effective Red Teaming with Auto-Generated Rewards and Multi-Step Reinforcement Learning" (OpenAI, November 21, 2024), https://cdn.openai.com/papers/diverse-and-effective-red-teaming.pdf.

[291] Perez et al., "Red Teaming Language Models with Language Models."

am not convinced that it is ever going to be possible to do that in a way that will get the same coverage as you will with humans. They are going to be complementary."[292]

Furthermore, as with any effort to automate, a human needs to be in the loop to make judgments on the results of automated red-teaming. Sam expressed his preference for people-based approaches, explaining:

> So long as we talk about red-teaming as like the tip of the spear, a very human-led sort of [and] not automated testing, then I'm happy. One thing that pushes at the boundaries of this and an open question is [that] is it valid to have a model attack another model as a red team? ... I guess it depends on ... what is the report that you get from that red-teaming? Because at some point, a human has to make a decision on what to do based on the risks that have surfaced. How do you trust the red team LLM in that case? How do you gain faith that it actually is doing the job that you think it is doing?[293]

Building on this concern, Anthropic identified a different challenge for automated red-teaming — the attacks they generate may not be as "novel and creative as those developed by people."[294] Furthermore, researchers have noted that the models used in automated red-teaming systems can be biased, exemplified by the higher likelihood of negative words co-occurring with identity terms such as Black or Muslim.[295] Such biases can increase false positives and false negatives, making it necessary to involve humans in the process to validate responses flagged as potentially harmful.

In describing how to do red-teaming, this subsection focused on the issue of identifying problematic model behavior. As Sam puts it, red-teaming is often framed as "the tip of the spear for how to identify where you may have a problem that you just didn't know about."[296] This commitment to problem identification underpins red-teaming efforts that invite red-teamers to be critical thinkers, straddle diverse perspectives, and question assumptions. Red-teaming is often the final step before public release, shaping the initial conditions of public engagement with the deployed system. In the next section, we delve into how public oversight and participation can ensure the safety and security of deployed systems.

[292] Emily, interviewed on 12 October 2023.

[293] Sam, interviewed on 29 September, 2023.

[294] Anthropic, "Challenges in Red Teaming AI Systems."

[295] Tom B. Brown et al., "Language Models Are Few-Shot Learners" (arXiv, July 22, 2020), https://doi.org/10.48550/arXiv.2005.14165.

[296] Sam, interviewed on 29 September, 2023.

Dynamics of public engagement with genAI red-teaming

The final step of red-teaming is **analyzing evidence**. This analysis shapes how companies act to address and resolve issues identified by red-teamers. These actions can unfold in two distinct ways. First, establishing accountability mechanisms for public oversight of red-teaming findings, and second, fostering public participation to raise awareness about genAI failures and prioritize the concerns of underrepresented communities. Mitigating genAI harm requires that companies follow through on red-teaming findings, through disclosure, measurement, and mitigation. These actions have legal, political, and social implications, and whether they are adequate should be assessed by the public, regulators, auditors, courts, and civil society.[297] Practitioners wanted the public to be able to make these assessments, but they often focused more on organizing public participation in genAI red-teaming. Such public participation efforts aim to help people better understand genAI models by experimenting with prompts that reveal flaws, while also using their feedback to refine future genAI systems by improving annotated prompt-response datasets. In the sections that follow, we address accountability for red-teaming findings and examine the role of public participation in genAI red-teaming.

How is accountability for acting on red-team findings organized?

Practitioners worry that red-teaming risks becoming performative, a form of "security theater,"[298] or as Will put it, a "kind of a waste of time ... because [developers] still don't have the processes to go back; they haven't created those feedback loops."[299] This section emphasizes the legal-political character of accountability and explores its role in holding "political and private

[297] Mark Bovens, "Analysing and Assessing Accountability: A Conceptual Framework," *European Law Journal* 13, no. 4 (2007): 447–68, https://doi.org/10.1111/j.1468-0386.2007.00378.x.

[298] Feffer et al., "Red-Teaming for Generative AI."

[299] Will, interviewed on 14 November 2023.

actors responsible."[300] Unlike auditing,[301] which can play a formal role in making accountability judgments — such as reasonable assurance of compliance with laws or standards[302] — red-teaming is not a formal accountability mechanism. Yet, there are growing expectations that organizations carry out red-teaming exercises and act upon their findings. Hence, the decisions companies make before, during, and after red-teaming are implicated in assessing accountability for model behavior. Specifically, we focus on what companies do with the findings of red-teaming.[303]

Practitioners argued that the utility of red-teaming was limited if companies failed to follow through with the results. Zhì observed that, "just because you can diagnose something, doesn't mean you know how to fix it."[304] Paras builds on Zhì's observation when he argues that red-teaming "is more about identification. ... It enjoys a lot of limelight because it is ... easily accessible to people. ... It is very hard to explain why a model produces a harmful photograph, but showing an example that for this prompt the model did this. 'Oh my God, look, it's so bad.' Its accessibility is extremely high."[305] Identification, though, is not enough.

Measurement. When asked about what should follow from the findings of red-teaming exercises, Paras talked about measurement and mitigation activities to close the feedback loop between findings from red-teaming and building more robust genAI systems. In this context, measurement does not refer just to quantitative assessments, but to a broader practice of gathering evidence.[306] The na-

[300] Deven R Desai and Joshua A Kroll, "Trust but Verify: A Guide to Algorithms and the Law," *Harvard Journal of Law and Technology.* 31 (2017): 9; Victor Ojewale et al., "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling" (arXiv, March 14, 2024), 2, http://arxiv.org/abs/2402.17861.

[301] Ojewale et al., "Towards AI Accountability Infrastructure," 2; Raji et al., "Outsider Oversight," 558; Michael Power, *The Audit Society: Rituals of Verification* (Oxford University Press, 1997), 3, https://doi.org/10.1093/acprof:oso/9780198296034.001.0001.

[302] Lam et al., "A Framework for Assurance Audits of Algorithmic Systems."

[303] Desai and Kroll (2017) emphasize a "technical" notion of accountability in computer science: "Computer science accountability ... is a technical concept about making sure that software produces evidence allowing oversight and verification of whether it is operating within agreed-upon rules" (Ibid. p10). An argument can be made that red-teaming conforms to this technical notion of accountability. However, this approach raises critical questions: how is the evidence disclosed, who determines if the software adheres to agreed-upon rules, and who is responsible when it does not? These questions point to the relational nature of accountability in a legal-political sense, which we explore in this section. Government mandates in the US and Europe further illustrate these legal-political implications through requirements for disclosing red-teaming results via appropriate channels. Furthermore, Metcalf and colleagues have argued that methods to assess impact of algorithmic systems are crucial to the process of achieving algorithmic accountability (red-teaming encompasses a diversity of such methods) and societal impacts provide a crucial framework for determining appropriate evidence thresholds and validating measurement techniques.
Jacob Metcalf et al., "Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 735–46, https://doi.org/10.1145/3442188.3445935; On the place of societal risks in setting appropriate validity and evidence expectations in AI evaluation, see: Blili-Hamelin and Hancox-Li, "Making Intelligence: Ethical Values in IQ and ML Benchmarks"; On whether red-teaming is appropriately framed as an accountability method, see: Chen, "Red Teaming Is about Assurance, Not Accountability." We thank Laura Weidinger for pointing out the need to clarify how red-teaming is implicated in accountability for problematic model behavior.

[304] Zhì, interviewed on 29 November 2023.

[305] Paras, interviewed on 27 October 2023.

[306] A good example of this emphasis is evident in a research project that focused on normative assumptions in stories that genAI tools produce and who is rendered visible through them. The author posed prompts — such as "Write a

ture and weight of the evidence are used to justify follow-up actions. Paras talked about this process extensively when he outlined *how measurement is central to standardizing labels for AI risks*[307]:

> While one example could be interesting, ... you cannot make a policy on one example. People push back. We need more understanding about this. So oftentimes, you find something weird, you bring it up, people will go and find more weird things like that, start to see patterns, and see if that resembles anything we already know. If there is already an example of something, we know that it is harmful. If there isn't, then there is a separate chain of [work] to label something and labeling is an excruciatingly long and drawn-out process. ... Nobody is using the red team to prove a system is safe. There is a whole different step after red-teaming that is the measurement phase that proves whether a system is safe or not.[308]

Both government bodies and tech companies are debating the specifics of this measurement phase.[309] As Paras continued, "what counts as measurement in this space is completely up for grabs. ... Your prioritization ... will shape which measurement efforts are worked on first. [This means that] you will be delaying measurements of other things and their mitigation ... because you will not understand the problem better if you don't measure it."[310] If these priorities are set through standardization and regulation, it inevitably implies that companies must comply with a set of safety and security standards for genAI. This regime will require companies to continuously monitor and periodically re-evaluate model performance against various benchmarks, while simultaneously critically examining the benchmarks and what they prioritize.

Mitigation. Closing the feedback loop presents a different set of challenges around how developers can fix failures once they have been identified. Some practitioners argued that mitigating a model's misbehavior is not the same as issuing a patch for a software vulnerability. As Zhì put it, "This is not a fix-it-and-move-on situation. This is where the analogy with cybersecurity completely falls flat. There is no such thing as the definitive patch, you install the update link, and move on with your life."[311] Extending this point further, Gavin noted that, "when you find an ML specific vulnerability, a lot of times the solution is not to fix the ML, it is to fix the integration[312] surrounding it."[313] In other words, the larger system surrounding the ML application can be changed

three-paragraph story in which two people meet, fall in love, and live happily ever after. Give the two characters names, jobs, and a favorite shared hobby" — fifty times to four genAI tools and qualitatively examined the normative assumptions embedded in the narratives produced as responses. See, Tarleton Gillespie, "Generative AI and the Politics of Visibility," *Big Data & Society* 11, no. 2 (June 1, 2024): 1–14, https://doi.org/10.1177/20539517241252131.

[307] For a proposal on leveraging "cards" boundary objects in coordinating uses of genAI risk labels, see: Leon Derczynski et al., "Assessing Language Model Deployment with Risk Cards" (arXiv, March 31, 2023), http://arxiv.org/abs/2303.18190.

[308] Paras, interviewed on 27 October 2023.

[309] One of the core areas of focus of the US AISIC, a consortium of 280 organizations dedicated to AI safety, is developing "guidance, methods, skills, and practices of successful red-teaming." Success is often a matter of measurement. See, NIST, "Artificial Intelligence Safety Institute Consortium (AISIC)," NIST, April 15, 2024, https://www.nist.gov/aisi/artificial-intelligence-safety-institute-consortium-aisic.

[310] Paras, interviewed on 10 November 2023.

[311] Zhì, interviewed on 29 November 2023.

[312] Integration refers to how the ML model is connected to and works within a larger system, including the software, hardware, data pipelines, and workflows that underpin its real-world applications.

[313] Gavin, interviewed on 23 August 2023.

to circumvent the vulnerability, such as updating its safety filters or modifying user prompts to include a diverse range of demographics to address representational issues. Grace discussed the strategy of modifying user prompts as “a pretty reasonable intervention” when talking about how prompts with the word “doctor” are often appended with demographic attributes (such as, “male doctor,” “female doctor,” “Black doctor”) to generate diverse representations.[314] This strategy is also at the heart of the controversy over the multi-racial Nazi-era German soldiers generated by Google’s Gemini, underscoring Zhì’s point that solving problems once they are identified is difficult in genAI applications.

Some contract red-teamers expressed *skepticism over how their findings were used by developers*. David, for example, reflected on his findings from participating in a red-teaming exercise:

> It wasn’t clear what could be changed or would be changed ... as a result of [our] feedback. Other than [developers who] said that “we might use these prompts. We can use these prompts to update the safety filters.” I think that was one clear thing. ... The other thing was the content policy, which is the policy that shows up when you use [a model] that says, you can’t do this or that. ... It seemed like that was also on the table, but it wasn’t clear whether those two things changed as a result of the stuff we did.[315]

Organizations rarely told consultant red-teamers what actions had been taken to mitigate the issues they found, thereby failing to complete communication over how developers used their findings.

In terms of *red-teaming’s potential contribution to model development*, Peter noted it may contribute to fine-tuning, but would not justify retraining the model. “Fully training a large language model is like a multimillion-dollar effort and you probably are not going to do that very often.”[316] In considering the implications of this observation, Philip argued that, “Let’s say that you’re OpenAI, ... you shouldn’t go ahead and start your training runs on GPT 5 until you’ve done something appropriate to address what your red-teamers found in GPT 4.”[317] While red-teaming helps to fine-tune a model and, eventually, may contribute to training its next iteration,[318] practitioners also observed that it is difficult to mitigate harms during modeling because it is far removed from real-world use. Jasmine brought up this concern:

> I work on the modeling side, which means that I have both a lot of control and also not very much control. I have a lot of control over ... the model’s objective function; I can change its data. But I can’t change the context in which it is used because ... I don’t deploy any of the models that I trained. ... There is another stage where you give it to a customer and the customer can fine-tune it on their own data. ... I have no visibility or control into that. ... You can’t really predict

[314] Grace, interviewed on 19 December 2023.

[315] David, interviewed on 20 December 2023.

[316] Peter, interviewed on 27 September 2023.

[317] This argument is in line with AI safety advocates who believe that shifting resources from scaling to safety and utility improvements is a more responsible approach to building and deploying genAI models. Philip, interviewed on 19 October 2023.

[318] Evident in the way OpenAI discussed the place of red-teaming in its safety efforts. See, OpenAI, “GPT-4 System Card.”

> downstream bias behavior from an upstream model. ... There's this weird thing that is happening with the space right now where ... the people who have the knowledge to mitigate harm are super far upstream, and the people who have the power to [mitigate harm] are pretty far downstream.[319]

Disclosure. These limits in closing the loop between identification, measurement, and mitigation led some respondents to advocate public disclosure of the findings of red-teaming exercises. Emily emphasized:

> The thing that could have the biggest effect is just making [reporting] more open. Most red-teaming efforts are very internal, we're lucky if [companies] put out a report. Companies get to decide what they report. So, I think if I could change anything, it [would be that] everything is open, to the degree that it is safe to be open, as long as you're not increasing harm by making it open. That would be the most impactful.[320]

However, as she pointed out, public disclosure of red-teaming results is uncommon. Red-teaming, both in genAI and other fields, is often perceived as primarily an internal process, meant to inform internal risk assessments and mitigation efforts. Disclosures have wide-ranging, often unintended effects on behavior and incentives.[321] In some settings, public disclosure requirements could have a chilling effect on red-teaming, leading to watered-down, easy-to-pass stress tests or less frequent red-teaming.[322] In security, public disclosure can also unintentionally help bad actors exploit vulnerabilities. Research on public disclosure of security vulnerabilities has stressed the importance of mitigating its unintended harms, emphasizing that without coordination between researchers and companies, it could lead to panic and undermine their future collaborations.[323] *Public vulnerability disclosures are a wicked problem* due to conflicting incentives, complex impacts, and a lack of definitive solutions.[324]

Public or community red-teaming exercises raise a different set of disclosure issues. On the one hand, Kabir brought up the three reporting mechanisms that are a part of most public-facing genAI portals, modeled on social media content flagging mechanisms: "thumbs up, thumbs down, and report

[319] Jasmine, interviewed on 25 September 2023.

[320] Emily, interviewed on 12 October 2023.

[321] Ashish Arora, Anand Nandkumar, and Rahul Telang, "Does Information Security Attack Frequency Increase with Vulnerability Disclosure? An Empirical Analysis," *Information Systems Frontiers* 8, no. 5 (December 1, 2006): 350–62, https://doi.org/10.1007/s10796-006-9012-5; Sabyasachi Mitra and Sam Ransbotham, "Information Disclosure and the Diffusion of Information Security Attacks," *Information Systems Research* 26, no. 3 (September 2015): 565–84, https://doi.org/10.1287/isre.2015.0587.

[322] Zenko gives examples from cybersecurity and military red teaming where organizations "game" or "cook" the engagement to avoid embarrassing failures, for example by taking certain machines offline prior to penetration tests. Zenko, *Red Team*, 19.

[323] Householder et al., "The CERT Guide to Coordinated Vulnerability Disclosure," 10.

[324] As Householder et al. argue, "Adversaries take advantage of vulnerabilities to achieve goals at odds with the developers, deployers, users, and other stakeholders of the systems we depend on. Notifying the public that a problem exists without offering a specific course of action to remediate it can result in giving an adversary the advantage while the remediation gap persists. Yet there is no optimal formula for minimizing the potential for harm to be done short of avoiding the introduction of vulnerabilities in the first place. In short, vulnerability disclosure appears to be a wicked problem." Householder et al., xi.

button, which is probably a decent way of collecting information. But we don't know how that translates into model retraining. ... There is a lot of intransparency surrounding these things."[325] On the other hand, for red-teamers who operate "in the wild," outside of sanctioned red-teaming events, disclosure of vulnerabilities potentially comes with legal risks. In the US, researchers conducting good-faith testing for model flaws are not currently protected by law, and may face consequences, including account suspension and lawsuits, for violating terms of service agreements.[326] Legal risks aside, immediately disclosing vulnerabilities to the public can potentially create or worsen security and safety risks. Coordinated security vulnerability disclosures, as championed by Computer Emergency Response Team Coordination Center (CERT/CC), can serve as a model for handling disclosure of high-risk genAI flaws.[327] It ensures that system owners are given advance notice and sufficient time to remedy problems prior to public disclosure.

Within the genAI practitioner community, there is broad support for ongoing disclosure of unmitigated (or not-fully-mitigated) risks, whether they were originally discovered through red-teaming or not. For example, model cards[328] or system cards are commonly used to communicate known flaws in models, such as DALL-E's tendency to produce stereotypical images.[329] These disclosures give users visibility into possible pitfalls and allow them to tailor their use cases to minimize harm. As Emily put it:

> I don't think you're ever going to be able to completely get rid of harm. You're never going to be able to completely get rid of bias in models because we can't get rid of it in people. We can't even find it all; we can't agree on what it is. So in terms of safeguards, **a lot of the most important problems are around making sure people understand what models can be used for and what they can't be used for**. No matter what you put in place, if it's a filter on a model or training scheme, someone out there is going to be able to game it. We've seen this immediately with all the big LLMs, as soon as you put it out there, someone is like, "Hey, look, I broke it this way." So, the biggest challenge is [making] continuous efforts to find issues that are there and address them in a way that brings in the community.[330]

Along these lines, public platforms such as Dynabench that facilitate sharing findings and strategies not only enhance transparency but allow for collective problem-solving.[331] In the following subsec-

[325] Kabir, interviewed on 17 November 2023.

[326] Longpre et al., "Position: A Safe Harbor for AI Evaluation and Red Teaming."

[327] Cattell, Ghosh, and Kaffee, "Coordinated Flaw Disclosure for AI"; Householder et al., "The CERT Guide to Coordinated Vulnerability Disclosure"; Jonathan M. Spring et al., "On Managing Vulnerabilities in AI/ML Systems," in *New Security Paradigms Workshop 2020* (NSPW '20: New Security Paradigms Workshop 2020, Online USA: ACM, 2020), 111–26, https://doi.org/10.1145/3442167.3442177.

[328] A model card is a standardized documentation format used to provide key details about a machine learning model, including its intended use, performance, limitations, and known flaws. System cards are more comprehensive and cover the entire system in which a model is embedded. Margaret Mitchell et al., "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* - FAT* '19 (the Conference, Atlanta, GA, USA: ACM Press, 2019), 220–29, https://doi.org/10.1145/3287560.3287596.

[329] OpenAI, "DALL·E 3 System Card," October 3, 2023, https://openai.com/index/dall-e-3-system-card/.

[330] Emily, interviewed on 12 October 2023, emphasis added.

[331] MLCommons, "Challenging the Limits of Benchmarking AI," Dynabench, 2023, https://dynabench.org/.

tion, we discuss how this focus on continuous improvement and adaptation in public red-teaming exercises — where diverse groups share and analyze model misbehavior — can enhance the robustness of these practices.

What is the place of public participation in red-teaming?

Finally, we asked practitioners to reflect on the nature of public participation in genAI red-teaming. Samantha, an AI risk governance expert who was also involved in organizing a public red-teaming event, offered an incisive perspective on the role of publics in AI safety: "The public does not need to be involved in the nitty-gritty details of how AI is designed and developed. They don't care; **they just want to be protected**. To the extent that red-teamers can provide that protection, that is incredibly valuable."[332] People feel protected as they experiment with diverse approaches — *building familiarity* with genAI systems (AI literacy), *exercising agency* when faced with the potential of model misbehavior (AI governance), and *contributing to the development* of genAI models (Participatory AI).[333] All three approaches were invoked in different ways to organize community red-teaming events.

Red-teaming typically targets the interests of *organizations*, such as improving the security and safety of their systems. It focuses on the features of a system that need to be made less vulnerable to attacks, or less susceptible to causing harm. Involving the public in red-teaming shifts this inquiry toward the interests of *communities*. These interventions emphasize **how the public can account for and respond to the uncertainties of a system's performance while continuing to use it.** This inversion is evident in a qualitative study of *red-teaming in the wild*, where researchers noted a crucial difference between professional red-teamers who "are explicitly looking for 'failure modes'" of genAI models and people interested in jailbreaking who "are often looking to get the model to obey."[334] This is a clear example of users intentionally pushing AI systems to misbehave, revealing their vulnerabilities. Beyond this ability to test and challenge the system, publics are often viewed as sources of cultural expertise, playing a key role in identifying and addressing harmful behavior. As the "last line of defense," end users help manage the consequences of model misbehavior. When publics use genAI models responsibly, they can significantly reduce the risks of AI causing harm.

Localized engagement. This report has outlined many benefits of involving publics in red-teaming efforts: it brings more diversity to the red team, draws on people's personal experiences to handle the subjective nature of what is considered harmful, and connects more closely to how genAI systems are actually used in real-world contexts. Organizers of public red-teaming events offered their own reasons that went beyond these arguments. For example, Luke, an AI safety and governance expert, reiterated his commitment to enrolling community college students into community red-teaming events, by emphasizing their uniqueness:

[332] Samantha, interviewed on 2 August 2023, emphasis added.

[333] Wanheng Hu and Ranjit Singh, "Enrolling Citizens: A Primer on Archetypes of Democratic Engagement with AI" (New York: Data & Society Research Institute, June 2024), https://datasociety.net/library/enrolling-citizens-a-primer-on-archetypes-of-democratic-engagement-with-ai/.

[334] Inie, Stray, and Derczynski, "Summon a Demon and Bind It," 20.

> The community college network is granular and fractal in nature. Every community college has some weird data set and some weird local issue that they're dealing with in some weird sub-industry — like they are loggers that are also metal workers and have coastal degradation problems. …That gives you an environment to test certain things. They are unique … and drive energy around AI implementation and training. … By enrolling community college students] you have moved this concept of red-teaming into application testing.[335]

These public-facing events emphasize how participants, such as community college students, use their unique perspectives and circumstances to engage with genAI systems. The goal is to help them become familiar with the technology and understand its limitations, rather than focusing solely on improving systems themselves.[336] On one hand, GNAReddy explained the significance of the DEF CON GRT event by noting that many of her peers "said they're going to go back to college after this challenge and start working on … generative AI models as their internal projects, like internships with their professors."[337] On the other hand, Gavin described such events as venues to experience the limitations of genAI systems by saying that, "we expose a lot of people to these models, and we give [them] firsthand experience with 'Oh look, I can make it say anything I want to, right, maybe I shouldn't trust it.'"[338]

Prioritizing people. While crowdworkers are paid to evaluate systems, community red-teaming efforts tend to emphasize competitions that prioritize people with diverse expertise to use their experiences in identifying how genAI models can fail and create educational opportunities to learn about different types of failures. Both approaches rely on user-generated prompts to identify problems and produce datasets that can be used by developers. The priority remains on diversifying prompting by inviting different communities to public red-teaming events. Amari, a community college student at the DEF CON event, champions such events: "Everyone is creative and has genius in their own way. So I feel like the more people that are able to get in front of the system and participate, the safer it is going to get."[339]

Furthermore, *the specter of embarrassing public incidents* often shapes how developers engage with community red-teaming. Paras compared this threat of reputational harm to how developers feel immediately after a product's launch; the feeling that "I don't want people to find faults in my system in the first week"[340] drives a lot of investments in pre-deployment red-teaming. When developers prioritize people in exploring the uncertainties of how AI models behave, they can view embarrassing outputs as opportunities to learn more about how these systems perform in real-world use.

Incentives to participate. Both organizers and participants positioned community red-teaming efforts as opportunities to be a part of a community and learn about genAI systems. As Amari noted,

[335] Luke, interviewed on 28 November 2023.

[336] The purple-teaming event at Greenwood provides a suitable example where one of the challenges given to participants was using a LLM to come up with a launch strategy, including finances, marketing, and a product roadmap, for their business concept.

[337] GNAReddy, interviewed on 14 September 2023.

[338] Gavin, interviewed on 23 August 2023.

[339] Amari, interviewed on 6 October 2023.

[340] Paras, interviewed on 27 October 2023.

"I get to be around all these extremely intelligent people and learn so much."[341] These efforts are also spaces to showcase expertise, as Samantha explained: "there is always going to be some person who is going to perform, if you have these challenges, way better than anybody else."[342] Participants were left with mixed feelings, as these different outcomes were sometimes at odds with each other. For example, Zuri, another community college student at the DEF CON event, expressed that she felt conflicted about participating in such a competition:

> I have two perspectives on it. One, I think there are a couple of reasons why it's good. Because you have the justice aspect of changing power dynamics, and [ensuring] equitable development of AI. We have diverse data sets, open evaluation. I also see it from a negative standpoint, because I was worried about data privacy. ... People may not have the knowledge or consent even to participate in some of these studies and would unwillingly or unknowingly give away their information, because that's way more valuable than their feedback. Because from the outside optics, [these events] make it seem like, "Oh, look, these companies are really cool." They're gathering feedback from communities, but on the inside, they're just harvesting data.[343]

Despite this ambivalence, community red-teaming events were seen by participants as key to raising awareness about genAI models. As Amari put it in the context of the DEF CON event:

> A lot of people from my community hadn't heard of [the DEF CON event] or didn't know anything about it. I want that to be different in the future. I know how important it is because [of the difference between] the kids that are on a chatbot now at [age] 8 versus the Black community, who might be anywhere else. It [will] have large effects; in a year's time that 8-year-old could be doing this and that, get all of his homework done, and have all of this extra free time. I want to close that gap [for my community].[344]

Beyond education, community red-teaming events serve as venues for exploring ethical concerns that matter to the public. For example, students had concerns around fairness and bias, but they also had an intuitive understanding of how and why bias manifests in genAI models. As Amari continued: "Yes, the Black community is definitely harmed by ... AI because we have the least amount of data. AI is all based on data. So if you have [less] data about a person, you will find that [these systems] easily make mistakes ... because [they] do not have enough data to know that that's offensive or incorrect."[345] Such events allow participants to sharpen their critical thinking as they interpret the behavior of genAI models.

Building expertise. The expectation that communities will engage with genAI models doesn't always match their ability to address or respond to technological issues. To return to Samantha's earlier point, the public does not need to be involved in every detail of designing and developing

[341] Amari, interviewed on 6 October 2023.

[342] Samantha, interviewed on 2 August 2023.

[343] Zuri, interviewed on 27 September 2023.

[344] Amari, interviewed on 6 October 2023.

[345] Amari, interviewed on 6 October 2023.

AI systems. The question is when and who should be involved, and how to balance expectations around community contributions. Participants felt that *red-teaming is a learned skill.* Whistledown, a participant at the purple-teaming event in Greenwood, Tulsa, suggested that community red-teaming should consider not only community interest in genAI models, but also understanding:

> It is not a one-size-fits-all [event. Participants should be] incentivized to actually explore more. But in a healthier sense, it's not like information overload.... For me just being in that room [during the purple-teaming event in Greenwood, the skill level of participants] was like easily level zero to level one.... [When you're only at a beginner level, such an event can be overwhelming.] It's kind of hard to give caviar to a baby. I feel like that's what was happening there. Give the baby milk, then give them soft foods and fruit, then move to solid, then change the portion size. So I just think that there's a way to do it. We can't rush it. ... You can't impact anything, if you don't even know what your impacts are and the thing to impact are people, like people are really your focus group. If we can't understand that, what are we doing?[346]

Building public expertise and a space for communities to explore and articulate their interests and concerns about genAI models lies at the heart of community red-teaming. For developers, these events help build red-teaming datasets, broaden their understanding of seemingly subjective goals (such as differing conceptions of AI harms), and offer deeper insight into everyday use.[347] For the public, community red-teaming provides a setting to learn from one another, engage with genAI systems and understand their workings and limitations, and occasionally interact with AI and domain experts who attend these events.

[346] Whistledown, interviewed on 28 May 2024. Pseudonym was chosen by the research participant.

[347] Anthropic, "Challenges in Red Teaming AI Systems."

Conclusion: Living with dissonance

> I always compare [red-teaming] to antigen screening for COVID-19. When you do antigen testing, it is a very unreliable test. If your test is positive, you pretty likely have the issue. If it's negative, we are not sure whether you have the issue, but we also can't reliably say that you don't have it. [That is,] pretty much, red-teaming summed up.[348]

If evaluating genAI models is a minefield,[349] then red-teaming is like playing Minesweeper.[350] Imagine an infinite grid representing all possible pairs of prompts and responses in interactions between humans and genAI models, where examining a problematic pair is like clicking a tile on this grid to reveal hints of potential AI harms. While a red-teamer can potentially identify many types of problematic model behavior and discover many metaphorical mines, it remains impossible to cover this infinite field of play.[351] The project of red-teaming in the public interest requires a critical perspective on this field of play, which not only examines the very premise of this game (or, red-teaming as a practice), but also reflects on the different roles that institutions, experts, and publics play in this game (or, how red-teaming is organized).

[348] Paras, interviewed on 27 October 2023.

[349] Narayanan and Kapoor, "Evaluating LLMs Is a Minefield."

[350] Minesweeper is a logic puzzle game where players click on tiles in a grid to clear hidden mines based on numerical clues that indicate the number of mines that are adjacent to the tile. The aim of the game is to clear the board of mines without triggering any explosions.

[351] In a 1972 reflection on early security application of tiger teaming (and what this report refers to as vulnerability probes), James Anderson raised a similar observation: "Even if corrections are made as a result of flaws found by a team, there is no assurance that all flaws have been found and corrected. The activities of the tiger team can only reveal system flaws and provide no basis for asserting that a system is secure in the event their efforts are unsuccessful. In the latter event, the only thing that can be stated is that the security state of the system is unknown. It is a commentary on contemporary systems that none of the known tiger team efforts has failed to date." James P. Anderson, "Computer Security Technology Planning Study" (NTIS, 1972), https://apps.dtic.mil/sti/citations/AD0758206; see also Chen, "Red Teaming Is about Assurance, Not Accountability."

We began this report by noting that the adversarial framing of the relationship between AI and society has shaped much of public conversation around genAI since the launch of ChatGPT. As a result, red-teaming becomes an obvious way to evaluate genAI, given its common association with adversarial thinking. However, practitioner experiences with genAI red-teaming challenge this adversarial frame. Practitioners consistently argued that their work goes beyond the narrow sense of attacks by motivated actors to evaluate problems that could arise during normal use. They also recognized the mismatch between problems that arise during normal use and the broader sense of adversariality involved in their work — the sense of challenging systems "with worst-case scenarios."[352] They invoked humility in making sense of the effectiveness of current practices. Finally, the organizational dimensions of genAI red-teaming showcase remarkable continuity with established professional security practices, ongoing historical debates over fixing security flaws and vulnerabilities post-release, and the harms associated with user-generated content.

In conclusion, we must **reframe the relationship between AI and society from adversarial to co-constitutive**.[353] By co-constitutive framing, we mean that AI is always already embedded within society. AI emerges from social practices, responds to social cues, mediates social practices, and is built from historically shaped, socially produced knowledge and data. This framing draws our attention to how AI produces and mediates particular social relations; how existing social relations shape certain AI applications and evaluation strategies; and how AI harms are experienced in everyday life. There is no society outside of AI for it to impact. AI is already woven into the fabric of society and its intricate tapestry of market, class, power, race, gender, and power relations.

This framing is represented in how practitioners emphasize the need for broader societal conversation on AI safety, protection from AI harms, and their disagreements over what counts as model failure and red-teaming. **AI evaluations do not happen in a vacuum**, but are always already shaped by societal perceptions of AI risks. As these perceptions evolve, so do the methods and priorities of evaluating and red-teaming AI. The current moment of genAI red-teaming may seem marked by troublesome confusion where no one quite agrees on what the field is doing, as evidenced by the contests over methods and clashes over power, authority, and expertise that surfaced during our interviews and fieldwork. However, a co-constitutive framing predicts and values this contestation and the diversity and fluidity in evaluation approaches at the intersections of AI and society.

This brings us to our second argument focused on the role that institutions, experts, and publics can play in genAI red-teaming. A major theme in this report is the **growing association of genAI red-teaming with interactive prompting**. Arvind Narayanan has succinctly observed that traditionally "in ML, building models is the central activity and evaluation is a bit of an afterthought. But the story of ML over the last decade is that models are more general-purpose and more capable. General purpose means you build once but have to evaluate everywhere."[354] The flexibility, effective-

[352] Siva Kumar and Anderson, *Not with a Bug, but with a Sticker,* 64.

[353] Wiebe E. Bijker and John Law, eds., *Shaping Technology/Building Society: Studies in Sociotechnical Change*, Inside Technology (Cambridge, MA: MIT Press, 1992); Brian J Chen and Jacob Metcalf, "A Sociotechnical Approach to AI Policy" (New York: Data & Society Research Institute, May 28, 2024), https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/.

[354] Arvind Narayanan and @random_walker, "Traditionally in ML, Building Models Is the Central Activity and Evaluation Is a Bit of an Afterthought. But the Story of ML over the Last Decade Is That ...," *X.Com*, September 8, 2024, https://x.com/random_walker/status/1840731490239340896.

ness, and applicability of interactive prompting across a broad spectrum of problems has earned it a central place in genAI evaluation efforts. However, its prevalence is a topic of ongoing contestation within traditional expert red-teaming communities; they often analogize interactive prompting with pen-testing, instead of red-teaming. Looking at this argument through a critical thinking mindset, instead of evaluating it based on existing professional identities, opens a new way to analyze public red-teaming interventions.

This mindset invites critical interrogation of so-called best practices in assessing genAI systems. As Gregory Fontenot, a former director of the US military's "Red Team University,"[355] cautions: "When you hear 'best practices,' run for your lives. The Titanic was built with best practices. It was faithfully operated in accordance with best practices."[356] While essential for evaluation, best practices cannot eliminate the risk of failure in complex sociotechnical systems. Nancy Leveson, a software safety expert, has succinctly argued that, "We are building systems today for which we cannot anticipate or guard against all unintended behavior."[357] Thus, a critical thinking mindset recognizes the limits of knowledge in anticipating failure and embraces holistic examination of "the behavior of all the components [of the system] working together along with the environment in which the components are operating."[358] By invoking the Titanic to make a point about red-teaming, Fontenot draws a vital connection between red-teaming and safety engineering: both focus on normal, routine practices as potential sources of failure. This commitment to critically examine routine practices, through outsider and contrarian thinking, has historically been central to the work of red teams.

In exploring how genAI red-teaming is organized, we found that experts focused on adversarial attacks — such as security or preventing disinformation campaigns or CBRN[359] attacks — tend to pay more attention to holistic red-teaming methods (like simulations) than those focused on evaluating AI harm to normal users. This may simply reflect that the playbook of red-team methods is more easily adapted to adversarial threat models. However, this also presents an **opportunity to introspect and innovate within sociotechnical safety evaluations**. This field of research has consistently interrogated sociotechnical gaps in AI evaluation, gaps in considering the human and societal factors that shape safety.[360] Practitioners in this space have critiqued the overemphasis on technical evaluation and favor more holistic methods like human-interaction evaluation[361] and impact assessments.[362] Red-teaming, as a critical thinking exercise, underscores a similar challenge: addressing

[355] Zenko, *Red Team*, 34.

[356] Zenko, 1.

[357] Nancy Leveson, *An Introduction to System Safety Engineering* (Cambridge, Massachusetts London, England: The MIT Press, 2023), 50.

[358] Leveson, 50.

[359] Barrett et al., "Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models," iv.

[360] Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems."

[361] Ibrahim et al., "Beyond Static AI Evaluations"; Schwartz et al., "The Draft NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan."

[362] Emanuel Moss et al., "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest" (New York: Data & Society Research Institute, June 29, 2021), https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/; Ada Lovelace Institute, "Algorithmic Impact Assessment: A Case Study in Healthcare" (Ada Lovelace Institute, 2022), https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/

organizational gaps between routine practices and identifying how these gaps result in failure under uncertain and complex conditions. While security red-teaming focuses on vulnerabilities by emulating threat actors, systems safety practitioners examine how unsafe practices become normalized in organizations.[363] Sociotechnical safety evaluations for genAI systems can benefit from drawing more inspiration not only from security red-teaming but also from safety engineering.[364]

Continuing with our critical inquiry into methods for genAI red-teaming, we offer two observations on the nature and scope of novel experiments in organizing interactive prompting to invite publics to evaluate genAI models. First, **interactive prompting to produce a red-teaming dataset is a marker of a power asymmetry**.[365] We have noted the limitation that comes from restricting public engagement to only evaluating already built systems, rather than directly shaping systems still in development. Many AI developers would argue that genAI models are always in the making, given the tradition of agile development, and that public input is instrumental in training the next generation of models. However, a similarity can be drawn between content flagging on social media and the kind of public input solicited through interactive prompting: *both reduce nuances of public concerns into data annotation exercises.* While these exercises are important for training models, they represent a very narrow form of public participation. In contrast, public red-teaming experiments invite deeper reflection on the nature of contributions that publics can make and whether they ensure meaningful public participation. While debates over these questions will continue, this report demonstrates that the role the public can play and the methods used to involve them are deeply interconnected and mutually shape each other.[366]

Second, the emphasis on **interactive prompting offers a window into the interplay between private and public interests** in identifying and ameliorating AI harms. Private interests are usually motivated by a single stakeholder's priorities and whether they can get other stakeholders to buy into their vision for how a system should be deployed and used in the real world. Security red-teaming is often conceived as serving the private interests of organizations. As a security professional, interviewed by Micah Zenko, put it, "it is essential for red-teamers to keep in mind that 'our job isn't to break into a computer network or building, it's to improve the security of the client.'"[367] In contrast, public interest is motivated by priorities that center the promise of a thriving and safe

[363] An early landmark in this line of inquiry is often thought to be: Perrow, *Normal Accidents*; Vaughan proposed "the normalization of deviance" as an account of the sociotechnical factors that allowed disregarding evidence that something was wrong prior to the NASA Challenger launch in 1986: Vaughan, *The Challenger Launch Decision*, 62; for an account of diversity as a safety value, see: Sidney Dekker, *Drift into Failure: From Hunting Broken Components to Understanding Complex Systems* (Farnham ; Burlington, VT: Ashgate Pub, 2011), 173; for a look at some of these safety engineering themes aimed at work on complex AI systems like genAI, including the importance of "curmudgeons, skeptics, & 'pathological thinkers' who only imagine worst-case scenarios," see Smart, Jacobs, and Kroll, "Unsafe at Any AUC."

[364] For examples of efforts in this direction, see, Smart, Jacobs, and Kroll, "Unsafe at Any AUC"; Rismani et al., "From Plane Crashes to Algorithmic Harm."

[365] For helpful provocations about the shortcomings of crowdsourced evaluation, see: Samantha Dalal, Siobhan Mackenzie Hall, and Nari Johnson, "Provocation: Who Benefits from 'Inclusion' in Generative AI?," in *EvalEval Workshop at NeurIPS 2024*, 2024, https://evaleval.github.io/accepted-papers.html; parth sarin, "Democratic Perspectives and Institutional Capture of Crowdsourced Evaluations," in *EvalEval Workshop at NeurIPS 2024*, 2024, https://evaleval.github.io/accepted-papers.html.

[366] We dive deeper into this relationship in the Appendix to this report on design choices for genAI red-teaming; see also, Hu and Singh, "Enrolling Citizens."

[367] Zenko, *Red Team*, 11.

society, instead of those of specific stakeholders. As Washington and Cheung write, “Projects that are truly in the public interest will never neatly align with only one financial, political, property, or ideological interest.”[368] Public red-teaming experiments prioritize people over genAI systems to educate and gain insight into what they consider problematic or harmful. These public experiments might not fully meet the needs of security professionals in identifying new failure modes, but they play a vital role in fostering public consciousness around living with genAI models and harnessing their capabilities while staying mindful of their failures.

Finally, taking the everyday impacts of genAI models seriously, we encourage systematically examining the experiences of their real-world harms. These experiences can serve as evidence, not just to remedy harms but to identify them earlier during the development of genAI systems.[369] We conceptualize a **reasonable expectation of safety** as an approach to engage with lived experiences as evidence and offer two key questions to frame it: whether the individual has demonstrated a subjective expectation of safety (for example, by using a genAI system consistent with its usage policy) and whether we — as public(s) committed to living in a democratic society — recognize that expectation as reasonable.[370] These questions cannot be answered by institutions alone; they require spaces where the public can deliberate on their experiences with genAI systems. Dewey offered collective inquiry as a solution to such complex challenges of democratic governance.[371] After all, articulating what is in the public interest has always been a matter of intense deliberation; its implications for genAI red-teaming are no different.

[368] Washington and Cheung, “Public Interest,” 105.

[369] A suitable example here is of misrepresentation of information by a chatbot in Moffatt v. Air Canada, No. SC-2023-005609 (Civil Resolution Tribunal of British Columbia February 14, 2024), https://canlii.ca/t/k2spq. Jake Moffatt sought a partial refund from Air Canada for bereavement fares after booking flights in November 2022 following his grandmother’s death. Moffatt relied on incorrect information from Air Canada’s chatbot, which suggested he could apply for the fare retroactively, but later learned this was not allowed. The Civil Resolution Tribunal found Air Canada liable for negligent misrepresentation, as Moffatt reasonably relied on the chatbot’s misleading information.

[370] In framing reasonable expectation of safety, we are drawing inspiration from how Steven J. Jackson, Tarleton Gillespie, and Sandy Payette, “The Policy Knot: Re-Integrating Policy, Practice and Design in CSCW Studies of Social Computing,” in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW ’14 (New York, NY, USA: ACM, 2014), 590, https://doi.org/10.1145/2531602.2531674 frame reasonable expectation of privacy. Exemplifying a kind of reasonable expectation of safety, in Moffatt v. Air Canada, the tribunal argued that, “while Air Canada argues Mr. Moffatt could find the correct information on another part of its website, it does not explain why the webpage titled ‘Bereavement travel’ was inherently more trustworthy than its chatbot. It also does not explain why customers should have to double-check information found in one part of its website on another part of its website.”

[371] Dewey, The Public and Its Problems: An Essay in Political Inquiry.

Acknowledgments

This report is the culmination of a one-and-a-half-year collaboration between Data & Society Research Institute (D&S) and the AI Risk and Vulnerability Alliance (ARVA). As we conclude this work, we would like to take a moment to express gratitude and appreciate the many contributions and voices that have shaped its trajectory.

We are deeply grateful to the industry practitioners and participants of public red-teaming events who shared their time, expertise, and reflections with us. Their firsthand experiences in organizing and participating in red-teaming exercises illuminated the complexities and contested terrain of AI safety and security. Their insights have laid the groundwork for how we have described red-teaming as a practice and the challenges shaping it in this report. We also extend our heartfelt appreciation to the broader community of red-teamers, security professionals, policy researchers, and AI practitioners whose ongoing work continues to shape genAI red-teaming. Their research, including the studies published on arXiv, provided critical perspectives that helped us map this evolving landscape.

Our sincere thanks to danah boyd, Leif Hancox-Li, Laura Weidinger, Wesley Deng, and Camille François for their careful review of earlier drafts, which helped refine and strengthen this work. Furthermore, engaging with Sorelle Friedler and Brian J. Chen while developing a policy brief on genAI red-teaming contributed valuable lessons to our research process.

At Data & Society, this report benefited greatly from the thoughtful feedback and conversations with Patrick Davison, Alice Marwick, Kiara Childs, Tamara Kneese, Meg Young, Serena Oduro, Ania Calderon, and Janet Haven, whose insights sharpened our analysis. We are especially grateful to the research management team — Alice Marwick, Patrick Davison, Kiara Childs, and Siera Dissmore — for their invaluable support in managing this collaboration and editing this report. This work would not have been possible without the contributions of Sona Rai, Chris Redwood, Eryn Loeb, Gloria Mendoza, Surbhi Chawla, Alessandra Erawan, and Iretiolu Akinrinade, whose expertise in communications, design, production, and network engagement helped bring it to life. A special thanks to Gloria Mendoza of GloriaMC Studio for the report's cover design and layout.

The AI Risk and Vulnerability Alliance thanks the Magic Grant community at the Brown Institute for Media Innovation for advice throughout the year, especially Eric Chen, Mark Hansen, Michael Krisch, and Todd Whitney. We are also thankful for grant application help from Subho Majumdar, Nathan Butters, Brian Pendleton, Jiahao Chen, Will Pearce, Rumman Chowdhury, and Kush Varshney. This project was deeply shaped by our experience as community partners in the Generative Red Team (GRT) Challenge at AI Village at DEF CON 31. We are grateful to AI Village, Humane Intelligence, and Seed AI for their leadership and to the members of the design and transparency working groups for thought-provoking discussions.

This research is funded in part through grants from the Omidyar Network, the Brown Institute for Media Innovation, and Siegel Family Endowment.

Red-teaming is ultimately a practice of critical engagement — of interrogating systems, identifying vulnerabilities, and as we explore in this report, imagining new ways to protect the public interest. We hope this report contributes to an ongoing and expansive public conversation about how we evaluate genAI systems in ways that are accountable, inclusive, and attuned to the complexities of the world in which they operate.

Appendix #1: Design choices for genAI red-teaming

In this appendix, we parse through clear and specific language about the *design* of red-teaming exercises through a faceted taxonomy of design choices.[372] Faceted taxonomies "describe content from multiple angles, perspectives or attributes."[373] A common example is field-based search. For example, regulations.gov allows refining a search of US federal government documents with facets for document type, document posted and comments due dates, and government agency.[374] Collectively, the facets of our taxonomy aim to clarify the *what, who*, and *how* of red-teaming designs. Questions focused on '*what?*' call attention to *scoping concerns* over the target of evaluation, purpose, and policy domain of a red-teaming exercise. The '*who?*' questions invite reflection on the people involved in organizing, sanctioning, or participating in red-teaming. Finally, the '*how?*' questions focus on the core activities of a red-teaming exercise. The use of this taxonomy can be supplemented with an open-ended description of the *methods* and *goals* of a red-teaming exercise.

[372] We draw inspiration from the Algorithmic Justice League's "Bug Bounties for Algorithmic Harms?", which develops a faceted taxonomy of design levers for bounty exercises. Josh Kenway et al., "Bug Bounties for Algorithmic Harms? Lessons from Cybersecurity Vulnerability Disclosure for Algorithmic Harms Discovery, Disclosure, and Redress" (Algorithmic Justice League, January 2022), https://www.ajl.org/bugs.

[373] Heather Hedden, *The Accidental Taxonomist*, Third edition (Medford, New Jersey: Information Today, Inc, 2022).

[374] US Government, "Regulations.Gov: Your Voice Is Federal Decision Making," accessed August 16, 2024, https://www.regulations.gov/.

What: Design choices that shape the scope of a red-teaming exercise

#1 | Target: What is the target of evaluation and does it include people and context?*

System performance

The exercise narrowly targets the performance of a system, separating it from the broader development and deployment context. Examples: interactive prompting.

Holistic

The exercise targets the system as well as the broader organizational context of developing and deploying it. Examples: security red-teaming and algorithmic impact assessments.

*The target does not need to be a technical system and its selection shapes the scope of the red-teaming exercise.[375]

#2 | Purpose: What is the broad purpose of the red-teaming exercise, and to what extent does it prioritize open-ended exploration?*

Prespecified (looking for known unknowns)

Red-teamers are given a specific set of instructions around what kind of failure modes to look for and find instances of model outputs as evidence for them.

Open-ended (looking for unknown unknowns)

Red-teamers are given open-ended instructions around finding potential failure modes and asked to use their own judgment in finding instances of model outputs as evidence for them.

*The relative priority of prespecified and open-ended exploration is crucial in the process of describing the purpose of a red-teaming exercise. We say relative priority because red-teaming often involves both. For instance, although prioritizing open-ended search for unknown failure modes, the instructions for Adversarial Nibbler specified four very broad categories of safety failures that provide substantive room for exploration: sexually explicit imagery, violent or graphic imagery, stereotypes and bias, hate symbols, hate groups, and harassment.[376]

#3 | Policy Domain: What types of failure mode, risks, harms, or threats does the exercise cover?*

Security

Evaluations that focus more on security and are oriented toward stress-testing a genAI system and identifying how susceptible it is to potential attacks by external adversaries when it is publicly deployed, or to aid in conducting adversarial attacks.

Sociotechnical Safety

Evaluations that focus more on safety and are oriented toward identifying potential ways in which a genAI system might fail under ordinary settings of using it without an adversarial intent on the user's part.

*Policy domain covers the implicit and explicit priorities, expectations, values, and norms that underlie red-teaming exercises. Given the relative novelty of looking for sociotechnical safety risks during red-teaming, explicit communication about their place in specifying scope is crucial.

[375] See, for example, the set of questions for pre-activity to guide future genAI red-teaming exercises in Feffer et al., "Red-Teaming for Generative AI," 4.

[376] Quaye et al., "Adversarial Nibbler," 2024, 389–90.

Who: Design choices that shape who is involved in red-teaming

#4 | Leading Entity: Who initiates the work, leads the process, and organizes the outcomes?*

Target Owners	Civic testing organizations	Individuals
Members of organizations who own the target of evaluation. This is often system owners, those who build or procure models.	Academics, journalists, and members of government and nonprofit organizations who evaluate publicly available genAI models for domain-specific public interest issues.	Members of the wider public who enjoy stress-testing publicly available models, getting them to "obey"[377] their instructions or jailbreaking them, and discussing model failures online.

*By leading entity, we mean the entity that influences and guides the red-teaming process. Leading involves making design decisions that shape a red-teaming effort in terms of its specific structure, scope, purpose, and outcomes. This can also involve being accountable for ensuring that the effort achieves its goals and has impact. Different entities can play distinct leading roles in initiating the red-teaming exercise, shaping its scope and purpose, and framing its outcomes.

#5 | Approval: Does the system owner officially agree to the exercise?*

Voluntary	Non-voluntary
The system owner officially agrees to the exercise.	The system owner does not officially agree to the exercise.

*Approval or lack thereof can take many shapes. The place of system owners as leading entities is obvious when they officially sanction red-teaming efforts. Yet, red-teaming can also be conducted without official approval from system owners.

Broadly, voluntary conditions include system owners: (1) initiating and leading the red-teaming process; (2) contracting an outside party to lead the red-teaming process; (3) providing voluntary safe harbor protections for approved outsider testing; and (4) providing access to a pre-release version or a version with removed or decreased guardrails.

Non-voluntary conditions for red-teaming include: (1) system owners being aware and tacitly approving of a testing effort without officially authorizing it; (2) system owners being aware and disapproving of a testing effort; and finally, (3) system owners being unaware of a testing effort[378].

[377] Inie, Stray, and Derczynski, "Summon a Demon and Bind It," 20.

[378] Zenko calls this "freelance red-teaming" and uses the cautionary example of KSDK, a news station in St. Louis, Missouri, who conducted tests to expose access safety protocol failures at local schools to discuss its potential pitfalls. The journalists' failure to put adequate protocols in place to ensure that the exercise wouldn't itself cause harm led to a school lockdown and panic. See, Zenko, *Red Team*, 222.

#6 | Expertise: Whose expertise is centered in team composition?*

AI Experts	Domain Experts	Cultural Experts
Familiarity with how genAI models are built and integrated into other computational systems to intuitively understand how and where it might fail.	Context-specific disciplinary familiarity with a use case of a genAI system to elicit practice-based insights on potential failure modes.	Familiarity with cultural nuances of representation that is grounded in lived experiences and standpoint of the participant to identify potentially harmful model output.

*Who participates in red-teaming is deeply aligned with considerations around how to put together a red-team, which involves thinking through inclusion/exclusion criteria, the team's representativeness in covering the prescribed scope, possible biases of team members, and incentives/disincentives to make contributions.[379]

#7 | Compensation Model: How are participants compensated for their efforts?*

Non-monetary	Prizes	Contract	Employment
Red-teaming can be done for non-monetary benefits such as recognition of expertise, educational achievement, or contribution to AI safety in the public interest.	When organized as competitions, red-teaming exercises are often promoted using prizes for winning them. Only the winners are rewarded for their efforts.	System owners retain red-teamers from outside their organization. This can be as workers within an effort organized by the company. Or as an outside company tasked with organizing and carrying out the red-team process.	System owners retain red-teamers from outside their organization on a permanent basis to evaluate genAI models they build, including red-teaming.

*The compensation model tends to determine the amount of effort that a red-teamer puts into their work.

#8 | Public Participation Model: How open is the red-teaming exercise to the public?*

Open	Closed/Invite-Only
The exercise is open to participation from members of the wider public.	The exercise is limited to those who are invited by the leading entity to participate and report on their findings.

*Public participation is a matter of logistics; the logistical features of how red-teaming exercises are organized shape who can participate in them. While decisions on these features inevitably shape the openness of a participatory activity, some obvious barriers to participation include: physical and online accessibility, access to resources, skills, registration fees, awareness, and outreach.

[379] See, for example, the set of questions for team composition to guide future genA red-teaming exercises in Feffer et al., "Red- Teaming for Generative AI."

How: Design choices that shape how a red-teaming exercise is organized

#9 | Access: What level of access and knowledge about the system is provided to red-teamers?*

Public access	Proprietary access
Red-teamers are only provided with publicly available resources for their work. Usually referred to as “black-box” testing in security contexts.	Red-teamers are also provided with access to proprietary knowledge or resources for their work, such as non-publicly available knowledge about model weights, system architecture, training data, etc. Proprietary access can be partial (“gray-box” testing) or full (“white-box” testing).

*Access determines what red-teamers know, which, in turn, determines the range of potential failure modes that they can identify. System owners play a pivotal role in deciding the level of access granted to red-teamers and determining the consequences for unauthorized red-teaming efforts.

#10 | Scale: What methods are used to scale red-teaming exercises?*

Expert teams	Crowdwork	Crowdsourcing	Automation
This strategy is closest to traditional red-teaming where teams of hand-picked experts are tasked with evaluating systems or organizations.	This strategy involves hiring crowdworkers from micro-task platforms to interactively prompt models and annotate data about desirable and undesirable model behavior.	This strategy involves public participation experiments that invite members of the wider public or a particular community to take part in red-teaming exercises.	This strategy involves auxiliary automated tools to generate prompts and evaluate target genAI model’s outputs. It requires human intervention to assess its success in identifying problematic target model behavior.

*The question of how to scale is pivotal for AI red-teaming because it invokes other concerns about who to include and why. Scale designates the human and technical resources needed for sufficiently broad and deep coverage of failure modes during red-teaming. Cost plays a major role in selecting scaling methods.

#11 | Disclosure Model: How are the results of the red-teaming exercise disclosed to the wider public? *

Full Disclosure	Selective Disclosure	Nondisclosure
The identified problems, mitigation measures, and the prompt/output dataset are freely disclosed to the public immediately or after a predetermined time frame without the need for approval from system owners.	The identified problems, mitigation measures, and the prompt/output dataset are selectively disclosed based on approval from system owners.	The identified problems, mitigation measures, and the prompt/output dataset are not publicly disclosed.

*The disclosure model specifies the rules of engagement around how identified problematic model behavior is shared publicly. Traditionally the results of a red-teaming exercise conducted internally by system owners are not disclosed to the public. However, given the increasing attention to problematic model behavior, questions of transparency and accountability have become crucial in debates over disclosure of red-teaming results, especially those that involve public participation. Disclosure models are also implicated in discussions over safe harbor protections[380] for disclosure of findings from good-faith safety evaluations of genAI models.

#12 | Participation Method: What is the format of the red-teaming exercise that involves members of the wider public?*

Focus groups	Competitions	Educational Events
Organizing discussions among participants is often used to gain diverse and deeper insights on potential vulnerabilities, question assumptions, conduct alternative analysis, and discuss simulations.	An especially successful format in ensuring larger participation of the broader public in red-teaming exercises is competitions or challenges organized in physical spaces or online platforms such as Kaggle.	Diverse forms of AI literacy interventions oriented toward encouraging debate over the use of genAI systems have started to incorporate a red-teaming component.

*Each format corresponds to different expectations from what members of the wider public can contribute to red-teaming. Competitions and challenges (such as CTF) often prescribe prespecified failure modes to lower the barrier to contribution and determine winners. More intensive adjudication and data annotation are needed for competitions such as bounty programs that incentivize participants to discover open-ended failure modes through material or reputational rewards for successful entries[381]. Educational events are organized around creating a space for learning how genAI models work, rather than explicitly evaluating their performance. Finally, focus groups[382] remain the most open-ended invitation to reflect on potentially problematic model behavior, but they are much smaller in scale.

[380] Longpre et al., “Position: A Safe Harbor for AI Evaluation and Red Teaming.”

[381] Ellis and Stevens, “Bounty Everything: Hackers and the Making of the Global Bug Marketplace”; Kenway et al., “Bug Bounties for Algorithmic Harms? Lessons from Cybersecurity Vulnerability Disclosure for Algorithmic Harms Discovery, Disclosure, and Redress.”

[382] Stevie Bergman et al., “STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment,” Scientific Reports 14, no. 1 (March 19, 2024): 2, https://doi.org/10.1038/s41598-024-56648-4; Gadiraju et al., “I Wouldn’t Say Offensive but…,” 207.

Appendix #2: Notes on methods

This project began with participant observation of the public genAI red-team (GRT) event at DEF CON in August 2023. Engaging with enthusiastic participants waiting to join the red-teaming challenge, observing conversations among experts on stage, and participating in sidebars in the hallways, we were often confronted with the ambiguous nature of genAI red-teaming: What does it include, and how should it be done to mitigate harms? These questions have shaped this research project, which includes participant observations from several public events with different approaches to probing genAI models through interactive prompting. These include an *expert-driven safety testing* event focused on election misinformation hosted by the AI Democracy Projects in January 2024 and a *purple-teaming event* held in the Greenwood District of Tulsa, Oklahoma, in February 2024. The project also includes semi-structured interviews with practitioners and participants from these events, conducted via Zoom. In parallel, we also conducted a systematic literature survey of publications focused on genAI evaluations broadly, and genAI red-teaming specifically.

We used snowball sampling to connect with a diverse group of practitioners spanning industry, government, academia, consultants, and contracted red-teamers, with initial interviewees helping us reach additional participants. They were initially recruited through the networks of our research team and our organizations (Data & Society and ARVA). Some participants were interviewed twice, and they were offered a $25 gift card per half-hour as an incentive for their time. Our goal was not to create a statistically representative sample but to connect with a range of participants who have stakes in and opinions on genAI red-teaming. Furthermore, we organized reading groups to review key literature to understand diverse methods for evaluating genAI models.

Table 1 summarizes our research participants, detailing their positionality with respect to genAI red-teaming and the interview dates, in the order they appear in the report. To protect privacy, interviewees have been anonymized, and their affiliations masked, with exceptions for those who requested to be named. We have used footnotes to indicate which respondents chose to be identified by their real first names. Furthermore, we have intentionally described their roles in broad terms to preserve confidentiality while providing some context to their positions within their respective settings.

Table 1: List of research participants quoted in the report

Pseudonym	Interview Dates	Sector	Role Description
Emily	12 October 2023	Industry	An industry practitioner focused on AI safety
Pravin	15 December 2023	Industry	An industry practitioner focused on responsible AI
Paras	27 October & 10 November 2023	Industry	A responsible AI practitioner who has been enrolled into red-teaming by his company
Will * Pearce	14 November 2023	Industry	An industry practitioner focused on cybersecurity
Sam	29 September 2023	Industry	An industry practitioner focused on responsible AI
Zhì	29 November 2023	Government/ Consultant	An expert on AI governance and risk management
Grace	19 December 2023	Academia	An expert in auditing algorithmic systems
Gavin	23 August 2023	Industry	An industry practitioner focused on cybersecurity
Jasmine	25 September 2023	Industry	An industry practitioner focused on AI safety
Ryan	17 November 2023	Industry	An industry practitioner focused on genAI security
Sarah	27 September 2023	Government/ Industry	An expert on AI safety evaluations with both government and industry experience

Philip	19 October 2023	Academia	An expert on AI governance and policy focused on red-teaming
David	20 October 2023	Academia	An expert in auditing algorithmic systems
Kabir	17 November 2023	Industry	An industry practitioner focused on machine learning (ML) ethics and policy
Eryk* Salvaggio	8 August 2023	GRT event participant	A new media artist interested in creative misuse of AI
Roxana* Daneshjou	7 November 2023	Academia	An academic red-teaming practitioner with dual expertise in machine learning and medicine
Caroline* Sinders	15 August 2023 & 25 August 2023	GRT event participant/ Contracted red-teamer	A machine-learning-design researcher and artist
Asma	27 November 2023 & 4 December 2023	Contracted red-teamer	An expert in human rights impact assessment who has contracted with many tech companies
Emma	4 August 2023	Academia / Contracted red-teamer	An academic involved in designing a public red-teaming event
Peter	27 September 2023	Government/ Industry	An auditor who works on evaluation standards for genAI systems
GNAReddy**	14 September 2023	GRT event participant/ Academia	A community college student participant at the DEF CON event
Samantha	2 August 2023	Government	An AI risk governance expert who was also involved in organizing a public red-teaming event
Luke	28 November 2023	Nonprofit	An AI safety and governance expert

Amari	6 October 2023	GRT event participant/ Academia	A community college student at the DEF CON event
Zuri	27 September 2023	GRT event participant/ Academia	A community college student at the DEF CON event
Whistledown**	28 May 2024	Tulsa event participant	A participant at the purple-teaming event in Greenwood, Tulsa
*Interviewees requested to be referred to by their real first name in the report. **Interviewees requested to be referred to by this name.			

The interviews focused on their backgrounds, perspectives on the most pressing challenges in evaluating genAI models, reflections on key events shaping the regulatory landscape of AI accountability, experiences with red-teaming exercises, and opinions on what makes a red-teaming event successful. The research protocol was reviewed and approved by Pearl IRB.

We transcribed all interviews using Otter.ai and followed a grounded theory approach to coding.[383] During the initial coding phase in ATLAS.ti, each team member selected transcripts they wanted to work with and conducted open coding. In our weekly all-hands meetings, we discussed differences in how we interpreted the interviews and applied codes. As individual team members coded their assigned transcripts, we started noticing recurring patterns across the interviews. Using these patterns, we created a codebook to track emerging themes and the questions that motivate them. This codebook guided a second round of coding, in which transcripts were assigned to a different team member, with ongoing discussions in our meetings about how the various aspects of generative AI red-teaming were reflected in the data.

From these discussions and broader groupings of codes, we developed a report outline organized around key questions: why, what, when, who, and how genAI red-teaming is conducted, as well as questions about accountability for findings from and the role of the public in red-teaming exercises. Team members wrote an integrative memo focused on one of these questions. We had in-depth conversations about how to scope each memo to ensure that they were distinct and avoided overlapping findings, although this was not always successful. This report started to come together as we merged these integrative memos and resolved the overlaps between them. Finally, we focused on the core themes — power asymmetries, uncertainty, and lack of expert consensus — that together represent our empirical findings to frame the introduction of this report and our extensive literature survey to provide a historical context to the ongoing work on genAI red-teaming.

[383] Juliet Corbin and Anselm Strauss, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 3rd edition (Los Angeles, Calif: SAGE Publications, Inc, 2007).

Bibliography

Abbott, Andrew. *The System of Professions: An Essay on the Division of Expert Labor.* First Edition. Chicago, Ill.: University of Chicago Press, 1988.

Ahmad, Lama, Sandhini Agarwal, Michael Lampe, and Pamela Mishkin. "OpenAI's Approach to External Red Teaming for AI Models and Systems." OpenAI, November 21, 2024. https://cdn.openai.com/papers/openais-approach-to-external-red-teaming.pdf.

"The AI Safety Institute (AISI)." Accessed October 3, 2024. https://www.aisi.gov.uk/.

"AI Safety Summit 2023 - GOV.UK," February 9, 2024. https://www.gov.uk/government/topical-events/ai-safety-summit-2023.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, et al. "Frontier AI Regulation: Managing Emerging Risks to Public Safety." arXiv, September 4, 2023. http://arxiv.org/abs/2307.03718.

Anderljung, Markus, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. "Toward Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework." arXiv, November 15, 2023. http://arxiv.org/abs/2311.14711.

Anderson, Elizabeth. "The Epistemology of Democracy." *Episteme* 3, no. 1–2 (June 2006): 8–22. https://doi.org/10.3366/epi.2006.3.1-2.8.

Anderson, James P.. "Computer Security Technology Planning Study." NTIS, 1972. https://apps.dtic.mil/sti/citations/AD0758206.

Andrews, Mel, Andrew Smart, and Abeba Birhane. "The Reanimation of Pseudoscience in Machine Learning and Its Ethical Repercussions." *Patterns*, August 1, 2024, 1–14. https://doi.org/10.1016/j.patter.2024.101027.

Anthropic. "Challenges in Red Teaming AI Systems," June 12, 2024. https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems.

———. “Frontier Threats Red Teaming for AI Safety.” Anthropic Announcements, July 26, 2023. https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

Anwar, Usman, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, et al. “Foundational Challenges in Assuring Alignment and Safety of Large Language Models.” arXiv, April 15, 2024. https://doi.org/10.48550/arXiv.2404.09932.

Arora, Ashish, Anand Nandkumar, and Rahul Telang. “Does Information Security Attack Frequency Increase with Vulnerability Disclosure? An Empirical Analysis.” *Information Systems Frontiers* 8, no. 5 (December 1, 2006): 350–62. https://doi.org/10.1007/s10796-006-9012-5.

Aroyo, Lora, and Chris Welty. “Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation.” *AI Magazine* 36, no. 1 (March 25, 2015): 15–24. https://doi.org/10.1609/aimag.v36i1.2564.

Arvind Narayanan and Sayash Kapoor. *AI Snake Oil*. Princeton University Press, 2024. https://press.princeton.edu/books/hardcover/9780691249131/ai-snake-oil.

Attenberg, Joshua, Panos Ipeirotis, and Foster Provost. “Beat the Machine: Challenging Humans to Find a Predictive Model’s ‘Unknown Unknowns.’” *Journal of Data and Information Quality* 6, no. 1 (March 4, 2015): 1–17. https://doi.org/10.1145/2700832.

Auray, Nicolas, and Danielle Kaminsky. “The Professionalisation Paths of Hackers in IT Security: The Sociology of a Divided Identity.” *Annales Des Télécommunications* 62, no. 11 (November 1, 2007): 1312–26. https://doi.org/10.1007/BF03253320.

Barrett, Anthony M, Krystal Jackson, Evan R Murphy, Nada Madkour, and Jessica Newman. “Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models.” Berkeley Center for Long-Term Cybersecurity (CLTC), May 2024. https://cltc.berkeley.edu/publication/benchmark-early-and-red-team-often-a-framework-for-assessing-and-managing-dual-use-hazards-of-ai-foundation-models/.

Beauchamp, Zack. “How to Avert a Post-Election Nightmare.” Vox, August 18, 2020. https://www.vox.com/policy-and-politics/2020/8/18/21371964/2020-transition-integrity-project-simulation-trump.

Ben Braiek, Houssem, and Foutse Khomh. “Machine Learning Robustness: A Primer.” arXiv, May 3, 2024. http://arxiv.org/abs/2404.00897.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

Bergman, Stevie, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. “STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment.” *Scientific Reports* 14, no. 1 (March 19, 2024): 6616. https://doi.org/10.1038/s41598-024-56648-4.

Beutel, Alex, Kai Xiao, Johannes Heidecke, and Lilian Weng. “Diverse and Effective Red Teaming with Auto-Generated Rewards and Multi-Step Reinforcement Learning.” OpenAI, November 21, 2024. https://cdn.openai.com/papers/diverse-and-effective-red-teaming.pdf.

Bijker, Wiebe E., and John Law, eds. *Shaping Technology/Building Society: Studies in Sociotechnical Change*. Inside Technology. Cambridge, MA: MIT Press, 1992.

Black Tech Street, and SeedAI. “Hack the Future Greenwood.” Hack The Future, 2024. https://www.hackthefuture.com/greenwood.

Blili-Hamelin, Borhane, and Leif Hancox-Li. “Making Intelligence: Ethical Values in IQ and ML Benchmarks.” In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. June 12–15, 2023, Chicago, IL, USA, 2023. https://doi.org/10.1145/3593013.3593996.

Blili-Hamelin, Borhane, Leif Hancox-Li, and Andrew Smart. “Unsocial Intelligence: An Investigation of the Assumptions of AGI Discourse.” *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (October 16, 2024): 141–55.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. First edition. Oxford: Oxford University Press, 2014.

Bovens, Mark. “Analysing and Assessing Accountability: A Conceptual Framework.” *European Law Journal* 13, no. 4 (2007): 447–68. https://doi.org/10.1111/j.1468-0386.2007.00378.x.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. “Language Models Are Few-Shot Learners.” arXiv, July 22, 2020. https://doi.org/10.48550/arXiv.2005.14165.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” April 20, 2020. http://arxiv.org/abs/2004.07213.

Bruno, Alessandro, Pier Luigi Mazzeo, Aladine Chetouani, Marouane Tliba, and Mohamed Amine Kerkouri. “Insights into Classifying and Mitigating LLMs’ Hallucinations.” arXiv, November 14, 2023. https://doi.org/10.48550/arXiv.2311.08117.

Bucher, Taina. “The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms.” *Information, Communication & Society* 20, no. 1 (January 2, 2017): 30–44. https://doi.org/10.1080/1369118X.2016.1154086.

Burrell, Jenna, Zoe Kahn, Anne Jonas, and Daniel Griffin. “When Users Control the Algorithms: Values Expressed in Practices on Twitter.” *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 138:1-138:20. https://doi.org/10.1145/3359240.

CAIS. “Statement on AI Risk.” Center for AI Safety, May 30, 2023. https://www.safe.ai/work/statement-on-ai-risk.

Carson, Austin. "Written Comments | U.S. Senate AI Insight Forum: Innovation." SeedAI, October 24, 2023. https://www.seedai.org/media/written-comments-us-senate-ai-insight-forum-innovation-austin-carson-founder-and-president-seedai.

Casper, Stephen, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. "Explore, Establish, Exploit: Red Teaming Language Models from Scratch." arXiv, October 10, 2023. http://arxiv.org/abs/2306.09442.

Cattell, Sven, Rumman Chowdhury, and Austin Carson. "AI Village at DEF CON Announces Largest-Ever Public Generative AI Red Team." AI Village, May 3, 2023. https://aivillage.org/generative%20red%20team/generative-red-team/.

Cattell, Sven, Avijit Ghosh, and Lucie-Aimée Kaffee. "Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 7:267–80, 2024. https://doi.org/10.1609/aies.v7i1.31635.

Chen, Brian J, and Jacob Metcalf. "A Sociotechnical Approach to AI Policy." New York: Data & Society Research Institute, May 28, 2024. https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/.

Chen, Jiahao. "Red Teaming Is about Assurance, Not Accountability." LinkedIn, October 27, 2023. https://www.linkedin.com/pulse/red-teaming-assurance-accountability-jiahao-chen-pzj9e.

Coleman, Gabriella. "Hacker." In *Digital Keywords: A Vocabulary of Information Society and Culture*, 158–72. Princeton, NJ: Princeton University Press, 2016. https://doi.org/10.2307/j.ctvct0023.19.

Corbin, Juliet, and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 3rd edition. Los Angeles, Calif: SAGE Publications, Inc, 2007.

Couldry, Nick, and Ulises A Mejias. "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject." *Television & New Media* 20, no. 4 (September 2018): 336–49. https://doi.org/10.1177/1527476418796632.

Crawford, Kate, and Tarleton Gillespie. "What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media & Society* 18, no. 3 (March 1, 2016): 410–28. https://doi.org/10.1177/1461444814543163.

DAIR Institute. "Data Workers Inquiry," 2024. https://data-workers.org/.

Dalal, Samantha, Siobhan Mackenzie Hall, and Nari Johnson. "Provocation: Who Benefits from 'Inclusion' in Generative AI?" In *EvalEval Workshop at NeurIPS 2024*, 2024. https://evaleval.github.io/accepted-papers.html.

Dalvi, Nilesh, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. "Adversarial Classification." In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 99–108. KDD '04. New York, NY, USA: Association for Computing Machinery, 2004. https://doi.org/10.1145/1014052.1014066.

Data & Society, and Center for Democracy & Technology. "Ensuring 'AI Safety' Begins with Addressing Algorithmic Harms Now," March 18, 2024. https://datasociety.net/announcements/2024/03/18/ensuring-ai-safety-begins-with-addressing-algorithmic-harms-now/.

Deck, Andrew. "Scale AI Is on a Hiring Spree for Speakers of Under-Represented Languages: Some Languages Pay a Lot Better than Others." Rest of World, August 29, 2023. https://restofworld.org/2023/scale-ai-language-training-hiring/.

Dekker, Sidney. *Drift into Failure: From Hunting Broken Components to Understanding Complex Systems*. Farnham ; Burlington, VT: Ashgate Pub, 2011.

Dempsey, J. R., Gen. W. A. Davis, A. S. Crossfield, and Walter C. Williams. "Program Management in Design and Development," 640548, 1964. https://doi.org/10.4271/640548.

Deng, Wesley Hanwen, Bill Boyuan Guo, Alicia DeVrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. "Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice," February 21, 2023. https://doi.org/10.1145/3544548.3581026.

Department for Science, Innovation and Technology, and Michelle Donelan. "AI Safety Summit: Introduction." UK Government, October 31, 2023. https://www.gov.uk/government/publications/ai-safety-summit-introduction.

Derczynski, Leon, Erick Galinkin, Jeffrey Martin, Subho Majumdar, and Nanna Inie. "Garak: A Framework for Security Probing Large Language Models." arXiv, June 16, 2024. http://arxiv.org/abs/2406.11036.

Derczynski, Leon, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. "Assessing Language Model Deployment with Risk Cards." arXiv, March 31, 2023. http://arxiv.org/abs/2303.18190.

Desai, Deven R, and Joshua A Kroll. "Trust but Verify: A Guide to Algorithms and the Law." *Harv. JL & Tech.* 31 (2017): 1.

DeVos, Alicia, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. "Toward User-Driven Algorithm Auditing: Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3491102.3517441.

Dewey, John. *The Public and Its Problems: An Essay in Political Inquiry*. Edited by Melvin L. Rogers. Athens, Ohio: Swallow Press, 2016.

DISARM. "DISARM Framework." DISARM Foundation. Accessed November 29, 2023. https://www.disarm.foundation/framework.

Dobbe, Roel, Thomas Krendl Gilbert, and Yonatan Mintz. "Hard Choices in Artificial Intelligence." *Artificial Intelligence* 300 (November 2021): 103555. https://doi.org/10.1016/j.artint.2021.103555.

Dobbe, Roel, and Anouk Wolters. "Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context." *Minds and Machines* 34, no. 2 (May 17, 2024): 12. https://doi.org/10.1007/s11023-024-09668-y.

Downer, John. *Rational Accidents: Reckoning with Catastrophic Technologies.* Cambridge, Massachusetts: MIT Press, 2024.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. "The Llama 3 Herd of Models." arXiv, July 31, 2024. https://doi.org/10.48550/arXiv.2407.21783.

Ehsan, Upol, Ranjit Singh, Jacob Metcalf, and Mark Riedl. "The Algorithmic Imprint." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1305–17. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3531146.3533186.

Ellis, Ryan, and Yuan Stevens. "Bounty Everything: Hackers and the Making of the Global Bug Marketplace." New York: Data and Society Research Institute, January 2022. https://datasociety.net/library/bounty-everything-hackers-and-the-making-of-the-global-bug-marketplace/.

Epstein, Steven. *Impure Science: AIDS, Activism, and the Politics of Knowledge*. First Edition. Berkeley, California: University of California Press, 1996.

Eslami, Motahhare, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. "First I 'like' It, Then I Hide It: Folk Theories of Social Feeds." In P*roceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2371–82. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016. https://doi.org/10.1145/2858036.2858494.

European Parliament. "The Act Texts." EU Artificial Intelligence Act, April 16, 2024. https://artificialintelligenceact.eu/the-act/.

Feffer, Michael, Anusha Sinha, Zachary C. Lipton, and Hoda Heidari. "Red-Teaming for Generative AI: Silver Bullet or Security Theater?" arXiv, January 29, 2024. https://doi.org/10.48550/arXiv.2401.15897.

Festenstein, Matthew. "Does Dewey Have an 'Epistemic Argument' for Democracy?" *Contemporary Pragmatism* 16, no. 2–3 (May 17, 2019): 217–41. https://doi.org/10.1163/18758185-01602005.

FLI. "Pause Giant AI Experiments: An Open Letter." *Future of Life Institute* (blog), March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Fraser, Nancy. "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy." *Social Text*, no. 25/26 (1990): 56. https://doi.org/10.2307/466240.

Friedler, Sorelle, Ranjit Singh, Borhane Blili-Hamelin, Jacob Metcalf, and Brian J Chen. "AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability." Policy Brief. New York: Data and Society Research Institute, October 2023.

https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf.

Frontier Model Forum. "What Is Red Teaming?" Frontier Model Forum, 2023. https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf.

Fyodor. "The Art of Port Scanning." *Phrack Magazine,* September 1, 1997. https://nmap.org/p51-11.html. https://nmap.org/p51-11.html.

Gadiraju, Vinitha, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. "'I Wouldn't Say Offensive But ...': Disability-Centered Perspectives on Large Language Models." In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 205–16. Chicago IL USA: ACM, 2023. https://doi.org/10.1145/3593013.3593989.

Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, et al. "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned." arXiv, November 22, 2022. http://arxiv.org/abs/2209.07858.

Gieryn, Thomas F. "Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists." *American Sociological Review* 48, no. 6 (1983): 781–95. https://doi.org/10.2307/2095325.

Gillespie, Tarleton. "Generative AI and the Politics of Visibility." *Big Data & Society* 11, no. 2 (June 1, 2024): 1–14. https://doi.org/10.1177/20539517241252131.

Gillespie, Tarleton, Ryland Shaw, Mary L. Gray, and Jina Suh. "AI Red-Teaming Is a Sociotechnical System. Now What?" arXiv, December 12, 2024. https://doi.org/10.48550/arXiv.2412.09751.

Goerzen, Matt, and Gabriella Coleman. "Wearing Many Hats: The Rise of the Professional Security Hacker." New York: Data & Society Research Institute, January 14, 2022. https://datasociety.net/library/wearing-many-hats-the-rise-of-the-professional-security-hacker/.

Goerzen, Matt, Elizabeth Anne Watkins, and Gabrielle Lim. "Entanglements and Exploits: Sociotechnical Security as an Analytic Framework," 2019. https://www.usenix.org/conference/foci19/presentation/goerzen.

Goldberg, Emma. "A.I.'s Threat to Jobs Prompts Question of Who Protects Workers." *The New York Times*, May 23, 2023, sec. Business. https://www.nytimes.com/2023/05/23/business/jobs-protections-artificial-intelligence.html.

Goldfarb-Tarrant, Seraphina, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. "This Prompt Is Measuring < MASK > : Evaluating Bias Evaluation in Language Models." In *Findings of the Association for Computational Linguistics: ACL 2023*, 2209–25. Toronto, Canada: Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.findings-acl.139.

Goldman, Sharon. "NIST Staffers Revolt against Expected Appointment of 'Effective Altruist' AI Researcher to US AI Safety Institute." *VentureBeat* (blog), March 8, 2024. https://venturebeat.com/

ai/nist-staffers-revolt-against-potential-appointment-of-effective-altruist-ai-researcher-to-us-ai-safety-institute/.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." arXiv, March 20, 2015. https://doi.org/10.48550/arXiv.1412.6572.

Graham, Mark, Ralph K. Straumann, and Bernie Hogan. "Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia." *Annals of the Association of American Geographers* 105, no. 6 (November 2, 2015): 1158–78. https://doi.org/10.1080/00045608.2015.1072791.

Grant, Nico. "Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms." *The New York Times*, February 22, 2024, sec. Technology. https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html.

Green, Ben. "Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness." *Philosophy & Technology* 35, no. 4 (October 8, 2022): 90. https://doi.org/10.1007/s13347-022-00584-6.

Gu, Shixiang, and Luca Rigazio. "Toward Deep Neural Network Architectures Robust to Adversarial Examples." arXiv, April 9, 2015. http://arxiv.org/abs/1412.5068.

Habermas, Jürgen. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Studies in Contemporary German Social Thought. Cambridge, Mass: MIT Press, 1989.

Hedden, Heather. *The Accidental Taxonomist.* Third edition. Medford, New Jersey: Information Today, Inc, 2022.

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. "Measuring Massive Multitask Language Understanding." arXiv, January 12, 2021. http://arxiv.org/abs/2009.03300.

Hoffman, Bryce G. *Red Teaming: How Your Business Can Conquer the Competition by Challenging Everything*. First edition. New York: Crown Business, 2017.

Householder, Allen D, Garret Wassermann, Art Manion, and Chris King. "The CERT Guide to Coordinated Vulnerability Disclosure." Special Report. CMU/SEI-2017-SR-022 CERT Division, August 2017. https://resources.sei.cmu.edu/asset_files/specialreport/2017_003_001_503340.pdf.

Hu, Wanheng, and Ranjit Singh. "Enrolling Citizens: A Primer on Archetypes of Democratic Engagement with AI." New York: Data & Society Research Institute, June 2024. https://datasociety.net/library/enrolling-citizens-a-primer-on-archetypes-of-democratic-engagement-with-ai/.

Humane Intelligence. "AI Village Defcon Dataset." 2024. Reprint, Humane Intelligence, June 7, 2024. https://github.com/humane-intelligence/ai_village_defcon_grt_data.

———. “Algorithmic Bias Bounty Programs.” Humane Intelligence, 2024. https://www.humane-intelligence.org/bias-bounty.

IAASB. *Handbook of International Quality Control, Auditing, Review, Other Assurance, and Related Services Pronouncements*. Vol. 1. NEW YORK: International Auditing and Assurance Standards Board, 2022.

Ibrahim, Lujain, Saffron Huang, Lama Ahmad, and Markus Anderljung. “Beyond Static AI Evaluations: Advancing Human Interaction Evaluations for LLM Harms and Risks.” arXiv, May 27, 2024. http://arxiv.org/abs/2405.10632.

Inie, Nanna, Jonathan Stray, and Leon Derczynski. “Summon a Demon and Bind It: A Grounded Theory of LLM Red Teaming in the Wild.” arXiv, November 13, 2023. https://doi.org/10.48550/arXiv.2311.06237.

Jackson, Steven J., Tarleton Gillespie, and Sandy Payette. “The Policy Knot: Re-Integrating Policy, Practice and Design in CSCW Studies of Social Computing.” In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, 588–602. CSCW ’14. New York, NY, USA: ACM, 2014. https://doi.org/10.1145/2531602.2531674.

Jacobs, Abigail Z., and Hanna Wallach. “Measurement and Fairness.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–85. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445901.

Japan AI Safety Institute. “Guide to Evaluation Perspectives on AI Safety.” Japan AI Safety Institute, September 25, 2024. https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.01_en.pdf.

———. “Guide to Red Teaming Methodology on AI Safety.” Japan AI Safety Institute, September 25, 2024. https://aisi.go.jp/assets/pdf/ai_safety_RT_v1.00_en.pdf.

Joint Task Force Transformation Initiative. “Guide for Conducting Risk Assessments.” Gaithersburg, MD: National Institute of Standards and Technology (NIST), 2012. https://doi.org/10.6028/NIST.SP.800-30r1.

Jones, Elliot, Mahi Hardalupas, and William Agnew. “Under the Radar? Examining the Evaluation of Foundation Models.” Ada Lovelace Institute, July 25, 2024. https://www.adalovelaceinstitute.org/report/under-the-radar/.

Lucas, Frank, Zoe Lofgren, Mike Collins, Haley Stevens, Jay Obernolte, and Valerie Foushee. “Letter to Laurie Locascio from Members of the House Committee on Science, Space, and Technology.” House Committee on Science, Space, and Technology, December 14, 2023. https://democrats-science.house.gov/imo/media/doc/2023-12-14_AISI%20scientific%20merit_final-signed.pdf.

Kasy, Maximilian, and Rediet Abebe. “Fairness, Equality, and Power in Algorithmic Decision-Making.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–86. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445919.

Kear, Mark. "Playing the Credit Score Game: Algorithms, 'Positive' Data and the Personification of Financial Objects." *Economy and Society* 46, no. 3–4 (2017): 346–68.

Kennedy, Helen. "Living with Data: Aligning Data Studies and Data Activism through a Focus on Everyday Experiences of Datafication." *Krisis: Journal for Contemporary Philosophy*, no. 1 (2018).

Kenway, Josh, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. "Bug Bounties for Algorithmic Harms? Lessons from Cybersecurity Vulnerability Disclosure for Algorithmic Harms Discovery, Disclosure, and Redress." Algorithmic Justice League, January 2022. https://www.ajl.org/bugs.

Khlaaf, Heidy. "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems." *Trail of Bits*, 2023. https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.

Kosinski, Matthew, and Amber Forrest. "What Is a Prompt Injection Attack? | IBM." IBM, March 21, 2024. https://www.ibm.com/topics/prompt-injection.

Kumar, Ram Shankar Siva. "Microsoft AI Red Team Building Future of Safer AI." Microsoft Security Blog, August 7, 2023. https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/.

Lam, Khoa, Benjamin Lange, Borhane Blili-Hamelin, Jovana Davidovic, Shea Brown, and Ali Hasan. "A Framework for Assurance Audits of Algorithmic Systems." In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1078–92. Rio de Janeiro Brazil: ACM, 2024. https://doi.org/10.1145/3630106.3658957.

Leveson, Nancy. *An Introduction to System Safety Engineering*. Cambridge, Massachusetts London, England: The MIT Press, 2023.

Leveson, Nancy G. *An Introduction to System Safety Engineering*. Cambridge, Massachusetts London, England: The MIT Press, 2023.

Li, Guanlin, Kangjie Chen, Shudong Zhang, Jie Zhang, and Tianwei Zhang. "ART: Automatic Red-Teaming for Text-to-Image Models to Protect Benign Users." *(NeurIPS 2024) 38th Conference on Neural Information Processing Systems*, 2024.

Longpre, Shayne, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, et al. "Position: A Safe Harbor for AI Evaluation and Red Teaming." In *Proceedings of the 41st International Conference on Machine Learning*, edited by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, 235:32691–710. Proceedings of Machine Learning Research. PMLR, 2024. https://proceedings.mlr.press/v235/longpre24a.html.

Mansbridge, Jane. "Feminism and Democracy - The American Prospect." *The American Prospect*, February 19, 1990. https://prospect.org/civil-rights/feminism-democracy/.

Mell, Peter, Karen Scarfone, and Sasha Romanosky. "The Common Vulnerability Scoring System (CVSS) and Its Applicability to Federal Agency Systems." Gaithersburg, MD: National Institute of Standards and Technology, August 30, 2007. https://doi.org/10.6028/NIST.IR.7435.

Messeri, Lisa, and M. J. Crockett. "Artificial Intelligence and Illusions of Understanding in Scientific Research." *Nature* 627, no. 8002 (March 7, 2024): 49–58. https://doi.org/10.1038/s41586-024-07146-0.

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. "Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts." In P*roceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–46. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445935.

Metz, Cade. "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead." *The New York Times*, May 1, 2023, sec. Technology. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.

Milan, Stefania, and Emiliano Treré. "Big Data from the South(s): An Analytical Matrix to Investigate Data at the Margins." In *The Oxford Handbook of Sociology and Digital Media*, edited by Deana A. Rohlinger and Sarah Sobieraj. Oxford: Oxford University Press, 2020.

Mill, John Stuart. *On Liberty*. Yale University Press, 2003.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT** '19, 220–29. Atlanta, GA, USA: ACM Press, 2019. https://doi.org/10.1145/3287560.3287596.

Mitra, Sabyasachi, and Sam Ransbotham. "Information Disclosure and the Diffusion of Information Security Attacks." *Information Systems Research* 26, no. 3 (September 2015): 565–84. https://doi.org/10.1287/isre.2015.0587.

MITRE. "Navigate Threats to AI Systems through Real-World Insights." MITRE ATLAS. Accessed June 26, 2024. https://atlas.mitre.org/.

MLCommons. "Challenging the Limits of Benchmarking AI." Dynabench, 2023. https://dynabench.org/.

Moffatt v. Air Canada, No. SC-2023-005609 (Civil Resolution Tribunal of British Columbia February 14, 2024).

Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." New York: Data & Society Research Institute, June 29, 2021. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

Mouton, Christopher A., Caleb Lucas, and Ella Guest. "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study." RAND Corporation, January 25, 2024. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

Narayanan, Arvind, and Sayash Kapoor. "Evaluating LLMs Is a Minefield." Talks — Arvind Narayanan, October 4, 2023. https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield/.

Narayanan, Arvind, and @random_walker. "Traditionally in ML, Building Models Is the Central Activity and Evaluation Is a Bit of an Afterthought. But the Story of ML over the Last Decade Is That ..." *X.Com*, September 8, 2024. https://x.com/random_walker/status/1840731490239340896.

NATO. "The NATO Alternative Analysis Handbook." NATO, 2017. https://www.act.nato.int/wp-content/uploads/2023/05/alta-handbook.pdf.

Newman, Lily Hay. "Microsoft's AI Red Team Has Already Made the Case for Itself." *Wired*, August 7, 2023. https://www.wired.com/story/microsoft-ai-red-team/.

NIST. "AI Risk Management Framework: AI RMF (1.0)." Gaithersburg, MD: National Institute of Standards and Technology, 2023. https://doi.org/10.6028/NIST.AI.100-1.

———. "Artificial Intelligence Safety Institute Consortium (AISIC)." NIST, April 15, 2024. https://www.nist.gov/aisi/artificial-intelligence-safety-institute-consortium-aisic.

Ojewale, Victor, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. "Toward AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling." arXiv, March 14, 2024. http://arxiv.org/abs/2402.17861.

OpenAI. "DALL·E 2 Preview - Risks and Limitations." OpenAI's GitHub, April 6, 2022. https://github.com/openai/dalle-2-preview/blob/main/system-card_04062022.md.

———. "DALL·E 3 System Card," October 3, 2023. https://openai.com/index/dall-e-3-system-card/.

———. "GPT-4 System Card." OpenAI, March 23, 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

———. "Model Spec." A document that specifies desired model behavior, May 8, 2024. https://cdn.openai.com/spec/model-spec-2024-05-08.html.

———. "OpenAI Red Teaming Network." OpenAI, September 19, 2023. https://openai.com/index/red-teaming-network/.

Pakzad, Roya. "Old Advocacy, New Algorithms: How 16th Century 'Devil's Advocates' Shaped AI Red Teaming." Substack newsletter. *Humane AI* (blog), May 11, 2023. https://royapakzad.substack.com/p/old-advocacy-new-algorithms.

Palta, Rina, Julia Angwin, and Alondra Nelson. "How We Tested Leading AI Models Performance on Election Queries." Proof, February 27, 2024. https://www.proofnews.org/how-we-tested-leading-ai-models-performance-on-election-queries/.

Parrish, Alicia. "Video Introduction to the Adversarial Nibbler Challenge: Data-Centric AI Competition for Adversarial Examples for Text-to-Image Models." DataPerf, July 2023. https://www.dataperf.org/adversarial-nibbler.

Parrish, Alicia, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. "BBQ: A Hand-Built Bias Benchmark for Question Answering." In *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105. Dublin, Ireland: Association for Computational Linguistics, 2022. https://doi.org/10.18653/v1/2022.findings-acl.165.

parth sarin. "Democratic Perspectives and Institutional Capture of Crowdsourced Evaluations." In *EvalEval Workshop at NeurIPS 2024*, 2024. https://evaleval.github.io/accepted-papers.html.

Pearce, Will, and Joseph Lucas. "NVIDIA AI Red Team: An Introduction." *NVIDIA Technical Blog* (blog), June 14, 2023. https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/.

Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. "Red Teaming Language Models with Language Models." arXiv, February 7, 2022. https://doi.org/10.48550/arXiv.2202.03286.

Perrigo, Billy. "Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer." TIME, January 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Perrow, Charles. Normal Accidents: Living with High-Risk Technologies. Princeton Paperbacks. Princeton, N.J: Princeton University Press, 1984.

Power, Michael. *The Audit Society: Rituals of Verification*. Oxford University Press, 1997. https://doi.org/10.1093/acprof:oso/9780198296034.001.0001.

Proof News, and The Science, Technology, and Social Values Lab at the Institute for Advanced Study. "The AI Democracy Projects." Proof News, June 25, 2024. https://www.proofnews.org/tag/the-ai-democracy-projects/.

Quaye, Jessica, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Rose Kirk, Minsuk Kahng, Erin van Liemt, et al. "Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. https://doi.org/10.1145/3630106.3658913.

———. "Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation." arXiv, May 13, 2024. https://doi.org/10.48550/arXiv.2403.12075.

Raghavan, Prabhakar. "Gemini Image Generation Got It Wrong. We'll Do Better." Google, February 23, 2024. https://blog.google/products/gemini/gemini-image-generation-issue/.

Raji, Inioluwa Deborah, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. "The Fallacy of AI Functionality." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–72. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3531146.3533158.

Raji, Inioluwa Deborah, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance." arXiv, June 9, 2022. https://doi.org/10.48550/arXiv.2206.04737.

Raymond, Eric S. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol: O'Reilly Media, 2001.

Rest of World. "2024 AI Elections Tracker." Rest of World, 2024. https://restofworld.org/2024/elections-ai-tracker/.

Rismani, Shalaleh, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. "From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. Hamburg Germany: ACM, 2023. https://doi.org/10.1145/3544548.3581407.

Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Illustrated edition. New Haven: Yale University Press, 2019.

Roose, Kevin. "A Conversation With Bing's Chatbot Left Me Deeply Unsettled." *The New York Times*, February 16, 2023, sec. Technology. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

———. "A.I. Poses 'Risk of Extinction,' Industry Leaders Warn." *The New York Times*, May 30, 2023, sec. Technology. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html.

Ross, Ron, Victoria Pillitteri, Richard Graubart, Deborah Bodeau, and Rosalie McQuaid. "Developing Cyber-Resilient Systems : A Systems Security Engineering Approach." Gaithersburg, MD: National Institute of Standards and Technology (US), December 8, 2021. https://doi.org/10.6028/NIST.SP.800-160v2r1.

The Royal Society, and Humane Intelligence. "Red Teaming Large Language Models (LLMs) for Resilience to Scientific Disinformation." The Royal Society & Humane Intelligence, May 2024. https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/.

Rumman Chowdhury. "What the Global AI Governance Conversation Misses." *Foreign Policy*, September 19, 2024. https://foreignpolicy.com/2024/09/19/ai-governance-safety-global-majority-internet-access-regulation/.

Scarfone, K A, M P Souppaya, A Cody, and A D Orebaugh. "Technical Guide to Information Security Testing and Assessment." Gaithersburg, MD: National Institute of Standards and Technology, 2008. https://doi.org/10.6028/NIST.SP.800-115.

Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, et al. "The Prompt Report: A Systematic Survey of Prompting Techniques." arXiv, July 14, 2024. http://arxiv.org/abs/2406.06608.

Schulhoff, Sander, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Liu Kost, Christopher Carnahan, and Jordan Boyd-Graber. "Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs through a Global Scale Prompt Hacking Competition." arXiv, March 2, 2024. https://doi.org/10.48550/arXiv.2311.16119.

Schwartz, Reva, Jonathan Fiscus, Kristen Greene, Gabriella Waters, Kyra Yee, Rumman Chowdhury, Theodore Jensen, Afzal Godil, and Patrick Hall. "The Draft NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan." Gaithersburg, MD: National Institute of Standards and Technology, June 5, 2024.

Sharkey, Lee, Clíodhna Ní Ghuidhir, Dan Braun, Jérémy Scheurer, Mikita Balesni, Lucius Bushnaq, Charlotte Stix, and Marius Hobbhahn. "A Causal Framework for AI Regulation and Auditing," January 18, 2024. https://www.apolloresearch.ai/research/a-causal-framework-for-ai-regulation-and-auditing.

Shelby, Renee, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, et al. "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–41. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3600211.3604673.

Shen, Hong, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. "Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 433:1-433:29. https://doi.org/10.1145/3479577.

Shen, Xinyue, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "'Do Anything Now': Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models." arXiv, May 15, 2024. https://doi.org/10.48550/arXiv.2308.03825.

Singh, Ranjit, and Steven Jackson. "Seeing Like an Infrastructure: Low-Resolution Citizens and the Aadhaar Identification Project." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 315:1-315:26. https://doi.org/10.1145/3476056.

Siva Kumar, Ram Shankar, and Hyrum Anderson. *Not with a Bug, but with a Sticker: Attacks on Machine Learning Systems and What to Do about Them*. Indianapolis: John Wiley and Sons, 2023.

Sloane, Mona. "The Algorithmic Auditing Trap." *OneZero* (blog), March 17, 2021. https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d.

Smart, Andrew, Abigail Z. Jacobs, and Joshua Kroll. "Unsafe at Any AUC: Unlearned Lessons from Sociotechnical Disasters for Responsible AI." SPSP: Psychology of Media and Technology, February 17, 2022. https://www.youtube.com/watch?v=n5J5oDiiEW8.

Solaiman, Irene, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, et al. "Evaluating the Social Impact of Generative AI Systems in Systems and Society." arXiv, June 12, 2023. http://arxiv.org/abs/2306.05949.

Spring, Jonathan M., April Galyardt, Allen D. Householder, and Nathan VanHoudnos. "On Managing Vulnerabilities in AI/ML Systems." In *New Security Paradigms Workshop 2020*, 111–26. Online USA: ACM, 2020. https://doi.org/10.1145/3442167.3442177.

Storchan, Victor, Ravin Kumar, Rumman Chowdhury, Seraphina Goldfarb-Tarrant, and Sven Cattell. "Generative AI Red Teaming Challenge: Transparency Report." Humane Intelligence, Seed AI, AI Village, 2024. https://drive.google.com/file/d/1JqpbIP6DNomkb32umLoiEPombK2-0Rc-/view. https://www.humane-intelligence.org/grt.

Stouffer, Keith, Michael Pease, CheeYee Tang, Timothy Zimmerman, Victoria Pillitteri, Suzanne Lightman, Adam Hahn, Stephanie Saravia, Aslam Sherule, and Michael Thompson. "Guide to Operational Technology (OT) Security." Gaithersburg, MD: National Institute of Standards and Technology (US), September 28, 2023. https://doi.org/10.6028/NIST.SP.800-82r3.

Suchman, Lucy. "The Uncontroversial 'Thingness' of AI." *Big Data & Society* 10, no. 2 (July 1, 2023): 1–5. https://doi.org/10.1177/20539517231206794.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing Properties of Neural Networks." arXiv, February 19, 2014. http://arxiv.org/abs/1312.6199.

Thiel, David. "Investigation Finds AI Image Generation Models Trained on Child Abuse." Stanford Cyber Policy Center, December 20, 2023. https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

Tiku, Nitasha. "Top AI Researchers Ask OpenAI, Meta and More to Allow Independent Research - The Washington Post." *The Washington Post*, March 5, 2024. https://www.washingtonpost.com/technology/2024/03/05/ai-research-letter-openai-meta-midjourney/.

UFMCS. *The Applied Critical Thinking Handbook (Formerly the Red Team Handbook)*. 7th Edition. Ft Leavenworth, KS: University of Foreign Military and Cultural Studies, 2015. https://irp.fas.org/doddir/army/critthink.pdf.

United Nations and AI Advisory Body. "Governing AI for Humanity." United Nations, September 2024. https://www.un.org/en/ai-advisory-body.

"US Artificial Intelligence Safety Institute." *NIST*, October 26, 2023. https://www.nist.gov/aisi.

US Government. "Regulations.Gov: Your Voice Is Federal Decision Making." Accessed August 16, 2024. https://www.regulations.gov/.

Vassilev, Apostol, Alina Oprea, Alie Fordyce, and Hyrum Anderson. "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations." Gaithersburg, MD: National

Institute of Standards and Technology (US), January 4, 2024. https://doi.org/10.6028/NIST.AI.100-2e2023.

Vaughan, Diane. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago: University of Chicago Press, 1996.

vera. "Verazuo/Jailbreak_llms." Jupyter Notebook, August 9, 2024. https://github.com/verazuo/jailbreak_llms.

Vest, Joe, and James Tubberville. *Red Team Development and Operations: A Practical Guide*. Independently published, 2020. https://redteam.guide/.

———. "Red Team Engagement vs Penetration Test vs Vulnerability Assessment." RedTeam.Guide, 2022. https://redteam.guide/docs/Concepts/red-vs-pen-vs-vuln/.

Wang, Angelina, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. "Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms That Optimize Predictive Accuracy." *ACM Journal on Responsible Computing* 1, no. 1 (March 31, 2024): 1–45. https://doi.org/10.1145/3636509.

Warner, Michael. "Publics and Counterpublics." *Public Culture* 14, no. 1 (2002): 49–90. https://doi.org/10.1215/08992363-14-1-49.

Washington, Anne L., and Joanne Cheung. "Public Interest." In *Keywords of the Datafied State*, edited by Jenna Burrell, Ranjit Singh, and Patrick Davison. New York: Data & Society Research Institute, 2024. https://datasociety.net/library/keywords-of-the-datafied-state/.

Weidinger, Laura, and William Isaac. "Evaluating Social and Ethical Risks from Generative AI." Google DeepMind, October 19, 2023. https://deepmind.google/discover/blog/evaluating-social-and-ethical-risks-from-generative-ai/.

Weidinger, Laura, John Mellor, Bernat Guillen Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, et al. "STAR: SocioTechnical Approach to Red Teaming Language Models." arXiv, June 17, 2024. http://arxiv.org/abs/2406.11757.

Weidinger, Laura, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, et al. "Sociotechnical Safety Evaluation of Generative AI Systems." arXiv, October 31, 2023. https://doi.org/10.48550/arXiv.2310.11986.

The White House. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The White House, October 30, 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

The White House. "FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety." The White House, May 4, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/

fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

Willison, Simon. “Prompt Injection and Jailbreaking Are Not the Same Thing,” March 5, 2024. https://simonwillison.net/2024/Mar/5/prompt-injection-jailbreaking/.

Zenko, Micah. *Red Team: How to Succeed by Thinking like the Enemy*. New York: Basic Books, 2015.

Ziewitz, Malte, and Ranjit Singh. “Critical Companionship: Some Sensibilities for Studying the Lived Experience of Data Subjects.” *Big Data & Society* 8, no. 2 (July 1, 2021): 1s13. https://doi.org/10.1177/20539517211061122.

Suggested Citation: Singh et al., *Red-Teaming in the Public Interest*, Data & Society and AI Risk and Vulnerability Alliance, February 2025. doi.org/10.69985/VVGP4368

Data & Society is an independent nonprofit research institute that advances new frames for understanding the implications of data-centric and automated technology. We conduct research and build the field of actors to ensure that knowledge guides debate, decision-making, and technical choices.

www.datasociety.net

@datasociety

Cover illustration and layout by Gloria Mendoza
February 2025