Harvard Data Science Review • Special Issue 4: Grappling With the Generative AI Revolution

Scaling Up Mischief: Red-Teaming AI and Distributing Governance

Jacob Metcalf[1] Ranjit Singh[2]

[1]AI on the Ground Program, Data & Society Research Institute, New York, New York, United States of America,

[2]Algorithmic Impact Methods Lab, Data & Society Research Institute, New York, New York, United States of America

Published on: Dec 13, 2023

DOI: https://doi.org/10.1162/99608f92.ff6335af

ABSTRACT

Red-teaming is an emergent strategy for governing large language models (LLMs), which borrows heavily from cybersecurity methods. Policymakers and developers alike have leaned heavily into this promising, yet largely unvalidated approach for regulating generative AI. We argue that AI red-teaming efforts address a particular and unique moderation need of LLM developers: scaling up human mischievousness by inviting a wide diversity of people to make the system misbehave in unsafe or dangerous ways. However, there are significant methodological challenges in connecting the practices of AI red-teaming to the broad range of AI harms that policymakers intend it to address. Caution is warranted as policymakers and developers invest significant resources into AI red-teaming.

Keywords: artificial intelligence, AI governance, red-teaming, large language models, generative AI, AI policy

Media Summary

Major figures in AI policy and industry have decided to emphasize a novel form of governance for AI systems called ‘red-teaming.’ This name is borrowed from cybersecurity, in which a ‘red team’ pretends to be an adversary and attempts to break into a system before deployment to make sure it is secure. For large language models (LLMs), such as ChatGPT, a red team attempts to make the system misbehave by outputting offensive, harmful, or dangerous content, the lessons from which are then used to patch the system. LLMs pose a peculiar moderation problem: their human interlocutors are globally diverse and highly mischievous, the models are ultimately statistically unpredictable, and yet developers are responsible for making their outputs safe. The White House has highlighted AI red-teaming as a central pillar of AI safety in its landmark voluntary agreement with AI developers, and in sponsoring the generative AI red-teaming event at the DEF CON hacker conference in Las Vegas in 2023. Yet this method is largely untested, and it is unclear that it can accomplish the outcomes that policymakers believe it can. We suggest that there are major methodological questions to be addressed before investing so much certainty in this governance strategy.

1. The Unique Challenge of Governing LLMs

Large language models (LLMs) pose a fundamental and unique governance problem: their behavior is stochastic, so when the public encounters these models through online portals, the relationship between input and output is not precisely predictable. Add the complexity of a chat interface organized around prompts and responses, and it becomes even more difficult to control the direction and intent of any conversation. Furthermore, the range of tasks that people might ask of these general purpose models—especially multimodal tools that can operate in text, code, and audiovisual formats—is so enormous that no group of human data scientists and moderators employed by one tech company could ever conceive of every potentially harmful,

dangerous, or offensive output that needed to be moderated in advance. As ethnomethodologist Harold Garfinkel noted in one of the early studies of everyday conversations (Garfinkel, 1967), humans are quite capable of applying a variety of interpretive frameworks to any interaction to meet their needs—there is no knowing or controlling in advance how a conversation will be received.

The plasticity of this human capacity to interpret responses irrespective of whether there is any logic to them means that in the context of conversations with LLMs, given that the field of prompts (inputs) cannot be controlled, efforts are directed toward controlling responses (outputs) such that their interpretation hopefully does not cause harm (Ganguli et al. 2022; Rajani et al., 2023). Most platform trust and safety challenges have heretofore have been focused on moderating and algorithmically serving user-generated content, whereas LLM developers need to attend to a unique multiparty *conversation at scale* between a stochastic machine and a globally diverse set of users. LLM platforms have, thus, created an emergent form of causal responsibility with which internet platforms grapple, and we should expect a variety of novel governance strategies to arise in response, for better or worse.

To put it simply, the challenge is: can we scale our governance strategies to match the scale at which human mischief happens on LLM applications? The rapid rise of 'AI red-teaming' as a leading prospective strategy for LLM governance is a response to this challenge. It is assumed that we could prefigure a range of possible prompts for which the LLM response is likely to be interpreted as harmful, and that this effort will be best served by gathering the right cohort of clever and diverse adversaries to challenge the systems.

2. The Emerging Governance Strategy of Red-Teaming LLMs

'Red teams' are a cybersecurity practice, wherein a team of experts adopts the behavior of a theoretical adversary and attempts to penetrate or otherwise break or disrupt a system prior to distribution (Zenko, 2015). The name derives from the tabletop war games that were used by the Pentagon and think tanks in WWII and the Cold War, in which the allies were represented by blue pieces and the team tasked with modeling potential behavior of adversaries were represented by red pieces (Pakzad, 2023). The purpose of red-teaming—in war games or cybersecurity—is to attempt to poke holes in a plan or a product, where the undesirable outcome—lose the war or get hacked—is clear to all parties.

The purpose and methods of red-teaming LLMs is (as of this time) much less clear. While the details differ, it generally consists of this: gather clever people, who may be the developer's employees, external consultants, or the general public. Give them access to either an unmoderated backend or public-facing version of the LLM portal. Invite them to prompt the system (i.e., provide written instructions to the model) in ways that might get it to misbehave or disclose undesirable features, and record the inputs and outputs (Dinan et al. 2019). The results are then ideally used to patch offensive, dangerous, harmful, or uncomfortably weird system behaviors, and in some cases are collected as part of a pre-deployment vulnerabilities report, such as OpenAI's GPT-4 system card (OpenAI, 2023).

There are AI security vulnerabilities that are amenable to traditional red-teaming techniques. But AI red-teaming appears to be an emergent discipline with distinct processes. Depending on where you sit, the adoption of the term ‘red-teaming’ to describe what is being done with LLMs is either an attempt to import legitimacy or a temporary waypoint on the path to likely new terminology. Amidst the terminological confusion we see a central epistemic challenge that has yet to be answered: what is the relationship between the activity, the data that is output from it, an actionable system vulnerability disclosed by such data, and potential harm to individuals or society created by the system?

Despite that confusion, the major actors in AI accountability have seized on this emergent practice as a central pillar in their plans for how we will ensure LLMs are safe and fair. For example, in the recent agreement between the White House and major generative AI developers about safety, red-teaming gets the pride of place as the first principle: “Commit to internal and external red-teaming of models or systems in areas including misuse, societal risks, and national security concerns, such as bio, cyber, and other safety areas” (White House, 2023). It is notable that those are ambitious outcomes far beyond the goals of cybersecurity red-teaming, a much more mature and clearly bounded practice. At the August 2023 Generative Red-Teaming event at DEF CON in Las Vegas, thousands of attendees at the notorious hacker conference were invited to participate in a gamified version of red-teaming sponsored by federal agencies, AI accountability nonprofits, and major AI developers (AI Red Team, 2023). Web-based tools like Lakera’s Gandalf role-playing game (Lakera.ai, 2023), and the forum Jailbreak Chat (Albert, 2023) that has users upvote clever prompts, similarly crowdsource a collection of prompts that produce harmful model outputs. Whether any of these are properly red-teaming is hotly disputed, but they are aligned around the core impulse of distributing key aspects of AI governance outward to a wider diversity of people. As one of the organizers of the DEF CON event said from the stage, “you get more interesting results if you use your own experience to attack the system” (Fieldnotes, August 10, 2023).

3. Conclusion: The Methodological Challenges of Red-Teaming for AI Governance

All signals point to policymakers and tech companies investing significantly more in red-teaming. We argue that the success of that effort depends on attending closely to methodological questions: What is red-teaming actually targeting? Can it effectively represent the societal risks that policymakers appear to have tasked it with? We now see free-ranging humans having ‘conversations’ with stochastic machines at scale for the first time. The space of what humans might ask a machine is practically unlimited, and so the mechanisms of moderation are restricted to limiting what the machine does in response. Given the weight that has been placed on red-teaming as an emergent governance strategy, the question is whether it produces the kind of data that illuminates a) a representative range of actual harms that LLMs produce, and b) actionable solutions to patch the systems such that those harms can be addressed.

Acknowledgments

We would like to acknowledge our research collaborators on the topic of AI red-teaming for their contributions to the ideas in this article: Borhane Blili-Hamelin, Beth Duckles, Emnet Tafesse, Tamara Kneese, Meg Young, Serena Oduro, Brian Chen, and Sorelle Friedler.

Disclosure Statement

This work is supported in part by the Open Society Foundation, the Omidyar Foundation, and the Siegel Family Foundation.

References

AI Red Team. (2023). *Hack the future*. https://www.airedteam.org/

Albert, A. (2023). *Jailbreak Chat.* https://www.jailbreakchat.com/

Dinan, E., Humeau, S., Chintagunta, B., & Weston, J. (2019). Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In K. Inui et al. (Eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4536–4545). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1461

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S, ... Clark, J. (2022). *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned*. ArXiv. http://arxiv.org/abs/2209.07858

Garfinkel, H. (1967). Common sense knowledge of social structures: The documentary method of interpretation in lay and professional fact finding. In *Studies in Ethnomethodology* (pp. 76–103). Prentice Hall.

Lakera.ai. (2023). Gandalf | Lakera – Test your prompting skills to make Gandalf reveal secret information. https://gandalf.lakera.ai/

OpenAI. (2023). *GPT-4 system card.* https://cdn.openai.com/papers/gpt-4-system-card.pdf

Pakzad, R. (2023, May 11). Old advocacy, new algorithms: How 16th century "devil's advocates" shaped AI red teaming. Substack newsletter. *Humane AI* (blog). https://royapakzad.substack.com/p/old-advocacy-new-algorithms

Rajani, N., Lambert, N., & Tunstall, L. (2023, February 24). *Red-teaming large language models*. https://huggingface.co/blog/red-teaming

White House. (2023, July 21). *Biden-Harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI.* https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

Zenko, M. (2015). *Red team: How to succeed by thinking like the enemy*. Basic Books.

References

- AI Red Team. (2023). *Hack the future*. https://www.airedteam.org/. https://www.airedteam.org/ ↩
- Albert, A. (2023). *Jailbreak Chat*. https://www.jailbreakchat.com/ ↩
- Dinan, E., Humeau, S., Chintagunta, B., & Weston, J. (2019). Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In K. Inui et al. (Eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4536–4545). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1461

  ↩
- Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S, ... Clark, J. (2022). *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned*. ArXiv. http://arxiv.org/abs/2209.07858

  ↩
- Garfinkel, H. (1967). Common Sense Knowledge of Social Structures: The Documentary Method of Interpretation in Lay and Professional Fact Finding. In *Studies in Ethnomethodology* (pp. 76–103). Prentice Hall. ↩
- Lakera.ai. (2023). *Gandalf | Lakera – Test your prompting skills to make Gandalf reveal secret information.* https://gandalf.lakera.ai/ ↩
- OpenAI. (2023). *GPT-4 system card.* https://cdn.openai.com/papers/gpt-4-system-card.pdf

  ↩
- Pakzad, R. (2023, May 11). Old advocacy, new algorithms: How 16th century "devil's advocates" shaped AI red teaming. Substack newsletter. *Humane AI* (blog). https://royapakzad.substack.com/p/old-advocacy-new-algorithms

  ↩

- Rajani, N., Lambert, N., & Tunstall, L. (2023). *Red-Teaming Large Language Models*. https://huggingface.co/blog/red-teaming ↩
- White House. (2023). *Biden-Harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI*. https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf ↩
- Zenko, M. (2015). *Red team: how to succeed by thinking like the enemy*. Basic Books. ↩