March 2024

Supporting NIST's Development of Guidelines on Red-teaming for Generative AI

Introduction

On October 30, 2023, President Biden released Executive Order 14110 (EO) pertaining to safe, secure, and trustworthy artificial intelligence (AI). The sprawling executive order sets the administration's priorities on various subjects related to the use of AI systems in everyday American life, ranging from establishing standards for safety to protecting American citizens' privacy. Specifically, the EO calls for NIST to "develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy… [including] rigorous standards for extensive red-team testing to ensure safety before public release." Given both CMU's ongoing collaboration with NIST on Artificial Intelligence issues, as well as our strong belief that experts must design, develop, and deploy AI systems responsibly to promote a more just and equitable society, the K&L Gates Initiative and the Block Center jointly hosted experts on campus from across the country in the public, private, and academic communities to support NIST's development of red-teaming guidelines.

In February 2024, the Block Center and the K&L Gates Initiative at CMU convened a workshop on CMU's campus to discuss red-teaming concerning Generative AI (GenAI). The convening consisted of several expert speakers and three panels focused on the following topics: (1) the frontiers of research on red-teaming of AI systems, (2) industry practices around AI red-teaming, and finally, (3) the policy and legal implications of AI red-teaming. This whitepaper synthesizes the key findings of the discussion during the day-long event.

Sponsors: The K&L Gates Initiative in Ethics and Computational Technologies at Carnegie Mellon University (CMU) aims to elucidate ethical and societal issues that arise in the development or use of computational technologies, including issues of fairness and justice, impact on individual autonomy and wellbeing, stakeholder participation and community empowerment, accountability, and governance, promoting benefits and mitigating risks and other related concerns. The Block Center's Responsible AI initiative brings together the university's cutting-edge educators and researchers and their expertise in partnership with public and private sector experts to advance effective policy-making and practical knowledge, generate thought leadership, and contribute to the timely discourse around the responsible use of AI.

NIST's Responsibilities under EO 14110

Before the three panel discussions began, Elham Tabassi of NIST presented to the convening on NIST's role as laid out by President Biden's AI EO. In February 2024, Elham was appointed Chief Technology Officer of the United States AI Safety Institute, responsible for leading key technical programs of the institute, focused on supporting the development and deployment of AI that is safe, secure and trustworthy.

Much of NIST's current activities as the coordinator for federal AI standards align with the President's expectations and roles within EO 14110. NIST collaborates closely with private sector industry and interested public sector communities to develop valid, scientifically rigorous methods, metrics, and standards for using AI systems. This collaboration is a multi-part process, including listening sessions, distillation of community feedback, creation of measurement standards, and providing support to stakeholders. These activities are ultimately meant to help advance the scientific underpinnings of guidelines in standards and then help to operationalize those guidelines for use by the American public.

NIST has undertaken a variety of activities in support of the EO's mission. In January 2023, NIST published its AI Risk Management Framework (AI RMF), a document developed alongside public and private partners to help organizations better manage and mitigate the risks associated with AI (NIST 2023). On February 8, 2024, the U.S. Department of Commerce announced the creation of the U.S. AI Safety Institute Consortium (AISIC). Housed under NIST, the Consortium will unite AI creators and users, academics, government and industry researchers, and civil society organizations to support developing and deploying safe and trustworthy AI. The AISIC currently includes over 200 member organizations from across the impacted community and is meant to help lead the United States Government in the science, practice, and policy of AI safety and trust. AISIC subcommittees will be responsible for assisting NIST in implementing a number of tasks outlined in President Biden's AI EO, including the development of a risk management framework specifically for GenAI systems, creating capability measurement guidelines for AI systems, helping to establish processes for identifying and labeling synthetic content generated by AI tools, as well as developing guidelines for red-teaming AI systems.

Background and Framing

President Biden's EO requires guidelines for AI red-teaming, which it defines as a **structured testing** effort, using **adversarial methods**, often in a **controlled environment** and in collaboration with

developers of AI, 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. With this in mind, Prof. Heidari argued that while the mention of a specific risk assessment method in a landmark policy document is a welcome development to many, significant questions remain about what a red-teaming exercise precisely entails, how it should be conducted to be effective and produce the desired outcomes, and subsequently, what role it can play in the future safety evaluation and regulation of GenAI.

To capture the complexity involved in red-teaming generative AI in practice, she mentioned recent work in her group analyzing publicly available reports on six recent red-teaming activities in the tech industry to evaluate generative AI models. She outlined their findings as follows:

1. Language models have been the primary objects of recent red-teaming evaluations (even though other forms of GenAI exist such as multimodal and text-to-image models).
2. The **threat models** and target **vulnerabilities** were often broad in nature (e.g., risks to national security or simply uncovering "harmful" model behavior). (It appears that this lack of specificity is meant to motivate exploring the entire AI's risk surface, but it can backfire and incentivize focus on easy-to-explore risks.)
3. **Team compositions** ranged from groups of subject matter experts to random samplings of community stakeholders to language models performing red-teaming (!). In some cases, red-teaming was conducted by internal teams prior to model release, while other red-teaming activities were conducted on publicly released models through APIs. The **resources** (including time, access level, and compute) available also varied based on team composition.
4. Red-teaming activities differed considerably in **processes** and **methods**. For example, some organizations chose to conduct a single round of red-teaming, while others saw red-teaming as an iterative process in which results from initial rounds of testing were used to prioritize risk areas for further investigation.
5. There is significant variation in the **publicly-shared outputs** of red-teaming efforts. In some cases, specific examples of risky model behavior uncovered were publicly shared. In other cases, findings were deemed “too sensitive” for publication.
6. Finally, the specifics of risk **mitigation** strategies were often not provided or evaluated.

In light of the lack of consensus around the scope, structure, and assessment criteria for AI red-teaming, she proposed a set of essential criteria that should be part of effective AI red-teaming guidelines, breaking them into Pre-activity, Within-activity, and Post-activity criteria.

Pre-activity criteria: Before the red-teaming exercise, it is essential to specify:

- What is the artifact under evaluation? Is it the AI model in isolation or the broader system in which it is to be embedded? Relevant factors here include the **version of the model** (including fine-tuning details), the safety **guardrails** in place, and **conditions of release**.

- What is the **threat model for the** red-teaming activity probes?
- What is the **specific vulnerability** it aims to find?
- What are the **criteria for assessing the success** of the red-teaming activity (including the **benchmarks** of comparison and **reproducibility** considerations)?
- What are the criteria for **team composition** and the inclusion/exclusion of members, and why? How many **internal vs. external members** belong to the team? What is the distribution of **subject-matter expertise**?

Within-activity criteria: During the activity, it needs to be elucidated:

- What resources are available to participants (including time and computing power)? Does it realistically mirror that of a potential adversary?
- What **instructions** are given to the participants to guide the activity? This can have important framing and priming effects.
- What kind of **access** do participants have to the model? (Some have argued black-box model access is insufficient for a rigorous evaluation.)
- What methods can members of the team utilize to test the artifact?
- What **auxiliary AI tools** (if any) are supporting the activity?

Post-activity criteria: After the activity, it is paramount to consider:

- **Reports and documentation** on the findings of the activity. Who will have access to those reports? Who can verify them? When and why?
- Whether the approach **"worked."** How **successful** was the activity in terms of the criteria specified pre-activity?
- A **blue-teaming activity**, crucial to proposing measures to mitigate identified risks. Is such an activity planned following red-teaming?

Heidari concluded by noting that while AI red-teaming is a potentially powerful method for risk identification and assessment, numerous factors can impact its outcomes and efficacy. It is, therefore, critical that red-teaming is *not* the sole focus of, nor a replacement for, a comprehensive program of risk management.

Panel 1: Forefronts of Red-teaming Research

The convening's first panel was moderated by Professor Zico Kolter, an associate professor at CMU's School of Computer Science. Professor Kolter was joined by Professor Graham Neubig – Associate Professor, CMU's School of Computer Science, Professor Sanmi Koyejo – Assistant Professor in Computer Science, Stanford University, and Professor Matt Frederickson – Associate Professor, CMU's School of Computer Science.

During the first panel, panelists highlighted how research in the field quickly transitions into practical applications, significantly impacting our daily lives. They emphasized the importance of actively engaging in red-teaming and jailbreaking large language models (LLMs) deployed in real-world scenarios, treating LLMs as integral software components within larger systems, and requiring specific **expertise** to assess their threat profiles. Moreover, they stressed the **dynamic nature of AI research**, demanding constant incorporation of new findings to ensure effective red-teaming.

The broader trustworthiness of AI systems, particularly in domains like healthcare and neuroscience, drew attention. Panelists discussed the complexity of societal systems and the necessity of considering potential harms and risks associated with AI technologies, especially in diverse demographic contexts.

The panel highlighted challenges in identifying when text generation systems malfunction and the importance of developing **frameworks for evaluation**. Panelists also raised various questions regarding the nature and scope of red-teaming, including whether it should involve an adversarial approach or focus on uncovering flaws within systems. They acknowledged the importance of considering worst-case scenarios and stressed the need for stress testing to push systems to their limits. Additionally, concerns were raised about the potential **psychological distress** caused by exposure to extreme content during testing. Furthermore, the discussion addressed challenges in integrating AI systems into larger frameworks and the necessity of rigorously testing across various contexts. The panelists emphasized the **importance of interdisciplinary research** and the need to consider socio-technical aspects of AI development.

Finally, the discussion underscored the importance of continued research into foundational AI models and their applications, considering both technical and societal implications. They highlighted the necessity of collaborative efforts and ongoing exploration to ensure the responsible deployment of AI technologies.

Panel 2: Industry Practices for Red-teaming

The convening's second panel was moderated by Professor Yonatan Bisk, an assistant professor at CMU's School of Computer Science. Professor Bisk was joined by Margaret Mitchell – Research and Chief Ethics Scientist at Hugging Face, Professor Zack Lipton – Assistant Professor of Machine Learning at Carnegie Mellon University and the Chief Scientific Officer of Abridge, and Ece Kamar – Managing Director of AI Frontiers at Microsoft Research.

There are diverse strategies and insights into the implementation of red-teaming practices in the field of AI and machine learning (AI/ML). One panelist described their organization's approach to red-teaming. By offering features such as thorough evaluation of AI models, detailed data analysis, and integration of user feedback, they have aimed to foster inclusivity in their red-teaming

endeavors. Notably, initiatives include providing **accessible evaluation interfaces for Subject Matter Experts (SMEs)** across diverse domains to **conduct adversarial tests on models within a production environment**. Additionally, they have simplified the red-teaming process by enabling low-code evaluation with just three lines of code. **Leveraging the leaderboard culture** inherent in traditional AI research, they have also successfully engaged developers through their red-teaming initiatives and provided report templates to facilitate participation.

Another panelist emphasized the importance of establishing clear legislation to **standardize** rigorous AI evaluation and red-teaming practices. This entails defining red-teaming and establishing frameworks to guide practitioners. Additionally, they stressed the significance of **direct engagement with stakeholders** to ensure responsiveness in AI development, particularly in highly-regulated domains.

The last panelist showcased their organization's integration of red-teaming across various stages and aspects of AI development, encompassing security vulnerabilities, privacy risks, and malicious use cases. Highlighting the importance of a red-teaming "platform" that extends beyond individual AI models, they provided examples of tools designed to aid practitioners in adversarial testing, including the Python Risk Identification Tool (PyRIT), an open source tool available on GitHub Hugging Face's Red Teaming Resistance Benchmark leaderboard.

Looking ahead, there is a consensus among experts that future red-teaming efforts should prioritize **empowering creative thinking** and exploring the **potential for human-AI collaboration**. Moreover, there's a recognized need to strike a balance between red-teaming for security vulnerabilities and for ensuring the effectiveness of AI applications.

There was emphasis on the need for clarity in defining red-teaming practices within the AI/ML community. There was a consensus among panelists regarding the necessity of **establishing best practices** for red-teaming, with a lean towards implementing red-teaming at the **system level in addition to solely focusing on individual models**. As the panelists advocated for breaking down tasks into different components with distinct focuses, there was a consensus among them on the importance of red-teaming as a crucial (but not only) component of responsible AI development. By fostering inclusivity, standardizing practices, and embracing diverse perspectives, the recommendations' aim is to ensure the ethical and effective deployment of AI technologies across various contexts.

Panel 3: Policy and Legal Implications of Red-teaming

The convening's third panel was moderated by Professor Hoda Heidari, the K&L Gates Career Development Assistant Professor in Ethics and Computational Technologies at CMU. Professor Heidari was joined by Katherine Lee – Senior Research Scientist at Google DeepMind, Lama Ahmad – Technical Program Manager for Policy Research at OpenAI, and Dean Ramayya Krishnan – Dean, Heinz College Of Information Systems and Public Policy and William W. and Ruth F. Cooper Professor Of Management Science and Information Systems at CMU.

Overall, the panelists argued that red-teaming necessitates multifaceted evaluation methods and mitigation strategies. Firstly, they stressed that policymakers and AI experts should collaborate to develop **crisp definitions** and **context-specific approaches** in the evaluation process. Once red-teaming is better defined, one can evaluate the model, system, or project against **measurable objectives to determine successes** and risks. **External experts** and stakeholders are considered crucial for comprehensive evaluation and assessment.

Further, panelists described how evaluation and mitigation of risks associated with AI are required at **multiple levels of granularity** and **different stages of the AI lifecycle**. **Memorization** in AI systems, for example, is a common risk that is brought up in conversations around mitigation. However, the harms are contextual, necessitating red-teaming efforts on both the system and its components. Another potential economic risk involves **algorithmic monoculture**. Panelists advocated for policy mechanisms to induce optimal algorithmic diversity while acknowledging the challenges associated with determining and enforcing optimal policies. **Risk tiering** and **field testing**, both pre- and post-deployment, were proposed as potential approaches. Regardless of the mitigation strategy employed, it should be tailored to the model, system, or project context and dependent on red-teaming insights (e.g., further **fine-tuning** versus **content policy modification**). Panelists also noted that efforts to mitigate AI risks are driven by **market forces**, global marketplace requirements (e.g., the EU's more restrictive risk tiering may influence foundation model developers to comply with such tiering to market globally), and **existing regulations** in various industries.

Lastly, the panel concluded by panelists noting that moving forward, enhanced disclosure and accountability policies will be crucial. Policy experts in the field recommend guidelines around disclosure, including suggestions for **dedicated disclosure processes** with the appropriate resources and responsibilities within companies, as well as establishing an organization similar to the Computer Emergency Response Team (CERT) for reporting AI issues. Clarity on **allocation of responsibilities** for various risks, from model development to downstream fine-tuning, is deemed essential for effective risk mitigation.

These takeaways underscore the complexity and importance of red-teaming in AI development and highlight the need for diverse evaluation approaches, mitigation strategies, and regulatory frameworks to ensure AI safety and security.

Conclusion

The key points from this expert convening can be summarized as follows:

- A **functional definition** of red-teaming, its components, scope and limitations, is necessary for effective red-teaming.
- GenAI research and practice communities must move toward **standards and best-practices** around red-teaming.
- The **composition of the red team** (in terms of diversity of backgrounds and expertise) is an important consideration.
- Red-teaming efforts should **address the broader system**—as opposed to individual components.
- The broader **political economy** (e.g., market forces, regulations) will influence the practice of red-teaming.