August 27, 2024

OpenPolicy's white paper

Innovative Leaders' Perspectives on AI Red Teaming

Overview- Definition, objectives and reach

AI red-teaming is a crucial practice for ensuring the safety, security, and reliability of AI systems. This practice involves structured testing efforts designed to identify and mitigate risks associated with AI technologies. Despite its central role in AI safety strategies, the application of AI red-teaming often lacks clarity and consistency, leading to significant variations in how it is defined and executed. Therefore, to maximize its effectiveness, there is a need for standardized guidelines, comprehensive evaluation frameworks, and transparent reporting of red-teaming activities and outcomes.

As defined in **Executive Order 14110**[1]**,** AI red teaming is "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 systems". Following **NIST AI RMF (AI 100-1),**[2] AI red-teaming falls under the Measures of accuracy which is defined by **ISO/IEC TS 5723:2022** as "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true." It is a structured effort using adversarial methods to find flaws and vulnerabilities in AI systems, typically conducted in a controlled environment and often in collaboration with AI developers. By adopting adversarial methods, red teams identify harmful outputs, unforeseen behaviors, and potential misuse risks, covering a wide range of issues that **extends beyond traditional cybersecurity**, including security, bias, misinformation, and harmful content generation applied to various AI models and applications, from chatbots in banking and healthcare to large language models (LLMs) like OpenAI's GPT-3 and Google's Bard.

The **CISA Roadmap for AI**[3] aligns closely with NIST's principles, reinforcing the role of AI red teaming in securing AI technologies and enhancing **overall resilience** by allowing for **proactive threat management**, **compliance, and risk mitigation** that strengthens the organization's ability to withstand and recover from adversarial attacks and other AI-related risks.

---

[1] See Executive Order 14110, under section 3 (Definitions)
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/
[2]See NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)
https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
[3]See CISA Roadmap for AI
https://www.cisa.gov/sites/default/files/2023-11/2023-2024_CISA-Roadmap-for-AI_508c.pdf

The **MITRE ATLAS framework**[4] provides a comprehensive overview of the tactics and techniques used against AI systems. Since the beginning of 2024, generative AI systems, such as Large Language Models (LLMs) and GPTs, have gained significant popularity. Although a definitive taxonomy of attacks on these systems has yet to be established, several types can be identified. Prompt injection is one of the most well-known attacks on LLMs, but many other techniques, such as indirect prompt injection and jailbreaking, also exist. Attackers may seek to generate illegal or copyrighted material, produce false or biased information, or leak sensitive data. If attackers can operationally compromise a model, they can introduce bias, linking cybersecurity directly to ethical concerns about fairness and integrity.

For example, a malicious actor with prejudiced motives might manipulate AI-driven loan applications to disadvantage certain demographic groups. They could potentially exploit the system either as an external threat actor or as an insider within the data science team. **This scenario underscores the insider risks and ethical implications that are inherently connected to cybersecurity. Similarly, sensitive issues such as healthcare decisions illustrate that security cannot be separated from other ethical and operational considerations.**

Thus, the reach of AI red-teaming includes diverse stakeholders, including AI developers, external experts, and automated tools, all working together to simulate potential attacks and evaluate the AI's robustness. Accordingly, the primary objectives of AI red-teaming are to identify vulnerabilities, such as biases and discriminatory outputs, and to develop strategies to mitigate these risks, aiming to **enhance the transparency and accountability of AI systems, ensuring their alignment with human and societal values.** Risk mapping to identify a broad spectrum of risks specific to the industry and application, using both automated and manual testing methods to uncover vulnerabilities and assess the model's performance in different scenarios, and continuous improvement through regular testing, refinement, and adaptation to evolving threats and model updates- are all part of this practices fundamentals objectives.

The objectives of security testing and red teaming, whether applied to AI systems or other technologies, are to identify vulnerabilities and improve the overall security posture. During a cybersecurity red team assessment, as broadly addressed in CISA Feb 28, 2023 **"CISA Red Team Shares Key Findings to Improve Monitoring and Hardening of Networks",**[5] the red team seeks to penetrate an organization's enterprise network to provoke a security response from the organization's personnel, processes, and technology. It also involves

[4]See MITRE ATLAS framework https://atlas.mitre.org/matrices/ATLAS
[5]See CISA Cybersecurity Advisory report on "CISA Red Team Shares Key Findings to Improve Monitoring and Hardening of Networks" https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-059a

collaboration with internal teams (Trust & Safety, Responsible AI, Security) and external evaluators to ensure comprehensive and objective risk assessments. By providing insights and recommendations, **AI red-teaming also guides policy and regulatory frameworks, helping to standardize AI evaluations**.

NIST AI 100-1 (RMF 1.0)[6] stresses that AI red-teaming is a critical component of managing AI risks and enhancing the resilience of organizations. By identifying and addressing potential flaws and vulnerabilities, red-teaming helps prevent data breaches, model theft, and the generation of harmful content and ensures the trustworthiness and reliability of AI systems, thereby supporting organizational resilience and overall security by reducing the risk of harmful or unintended consequences. **Continuous red teaming efforts,** automated and with human supervision, ensure that AI systems remain robust and effective against evolving threats, contributing to the organization's resilience against adversarial attacks in real-time, ensuring they meet high safety and reliability standards, which is crucial not only to enhance overall security but also for user confidence and regulatory compliance. Enhanced transparency and reporting build trust among users, stakeholders, and regulators, promoting the wider acceptance, trust and adoption of AI technologies.

For example, suppose an organization has developed an AI model and validated its performance against anticipated real-world inputs, achieving satisfactory outcomes. The organization may have even incorporated adversarial examples into the training data to enhance the model's resilience. While these measures are beneficial, red teaming goes further by rigorously evaluating the model's resistance to both established and emerging attacks through realistic adversarial simulations.

This practice is particularly crucial for generative AI deployments, given the unpredictable nature of their outputs. The ability to test for harmful or otherwise unwanted content is essential not only for maintaining safety and security but also for ensuring trust in these systems. Although numerous automated and open-source tools are available, they have limitations and cannot substitute for the in-depth analysis provided by comprehensive AI red teaming. Many of these tools are static prompt analyzers that rely on pre-written prompts, which defenses commonly block due to their pre-existing awareness. For tools that employ dynamic adversarial prompt generation, crafting effective system prompts can be challenging, and some prompts deemed "malicious" may not actually pose any harm.

Thus, while automated tools offer some value, in-depth **AI red teaming remains indispensable** for thoroughly evaluating and securing AI systems against evolving threats.

[6]See NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)
https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Structured red-teaming practices are essential for organizations to comply with emerging regulations and standards, mitigating legal, ethical and reputational risks. By integrating AI red teaming into safety protocols as best practices and standards, organizations can better navigate the complexities of AI risks and build more secure, resilient, and trustworthy AI systems. **Proactively managing the various vulnerabilities and potential misuse risks to better anticipate and counter potential threats, thereby strengthening resilience against adversarial attacks and other AI-related vulnerabilities.**

Secure and trustworthy AI red teaming use and implementation

Following the above and our partners' Red Teaming reports,[7] OpenPolicy encourages examining the following implementation recommendations to ensure secure and trustworthy AI red teaming methods as the AI regulatory framework evolves.

1. **Functional AI red teaming and testing:**

   As mentioned above, red teaming involves conducting adversarial testing to identify vulnerabilities by simulating threat actor behavior, including creating risky prompts and assessing model responses. To complement this, functional testing should employ systematic approaches with standardized methodologies, such as behavioral and keyword-based prompts, to evaluate the robustness of safeguarding mechanisms. Given the risks and unpredictable vulnerabilities of LLMs under AI red teaming, it is beneficial to **assess and integrate automation tools for specific testing elements.** These tools can generate semantic variations of prompts and perform initial risk scoring of outputs to ensure real-time risk mitigation and reduce the need for extensive resources.

   Integrating automated tools with the nuanced understanding of human testers in AI red teaming efforts is crucial. This approach would involve enlisting both general red teamers for broad flaw detection and specialists for pinpointing nuanced issues such as misinformation and deepfakes. By fostering a collaborative environment that draws on the collective wisdom of the community, organizations can achieve thorough evaluations and leverage varied skill sets. **Encouraging practices like bias bounties and vulnerability disclosures further helps uncover flaws that might otherwise go unnoticed.**

[7]HiddenLayer "Guide to AI Red Teaming" https://shorturl.at/xEWpF, HiddenLayer "Financial Case Study: AI Red Teaming" https://hiddenlayer.com/research/financial-case-study-ai-red-teaming/ and ActiveFence "Mastering GenAI Red Teaming: Insights from the Frontlines" report https://www.activefence.com/research/genai-red-teaming,

Manual efforts should be directed toward complex or creative tasks that necessitate human judgment and ingenuity. Automated AI red teaming tools can efficiently scale and promptly identify risks, thereby enhancing the overall security posture. This allows human experts to focus on areas where their skills are most needed, particularly where automated tools may be inadequate. Automated systems often lack the adaptive creativity and nuanced perspective that human red teamers provide, making human involvement crucial for comprehensive security assessments.

This is why there is **a need to address both automated and manual AI red teaming,** providing a holistic approach to defend, protect, and detect evolving threats and constantly changing adversarial behaviors. By combining the scalability and efficiency of automated tools with the expertise and creativity of human red teamers, organizations can ensure a robust defense against sophisticated attacks.

- **Real Life Case Study:** OpenPolicy's partner, HiddenLayer, reported[8] that their engagement with a client in the financial services sector highlighted the critical importance of red teaming. The client had an AI model designed to identify fraudulent transactions. During testing, HiddenLayer discovered various methods attackers could use to bypass the fraud models and crafted adversarial examples. By collaborating closely with the client, they identified examples with minimal feature modifications, which provided valuable guidance to the data science teams for retraining the models to be less vulnerable to such attacks.

  Had adversaries identified and exploited these weaknesses first, it could have resulted in significant financial losses. By uncovering these vulnerabilities early, the client was able to fortify their defenses and enhance the comprehensiveness of their models. This approach not only protected the institution's assets but also maintained an excellent customer experience, which is crucial for their success.

Therefore, Incorporating both automated and manual AI red teaming ensures that organizations can effectively adapt to and counter evolving threats, safeguarding their operations and maintaining trust with their customers. **This dual approach provides a comprehensive defense strategy that leverages the strengths of**

[8]See HiddenLayer Financial Services Case Study https://shorturl.at/mulen

both automated systems and human expertise, ensuring high safety and reliability standards for AI mechanisms and models.

2. **Emphasis on Third-Party Red Teaming:** While internal red teaming efforts can offer valuable insights, our clients' feedback indicates that relying solely on internal teams often results in less robust findings due to a lack of domain-specific expertise, which only emphasizes the ecosystem-growing consensus that regulation, standards, and best practices should underscore the use of third-party red teaming. External teams tend to uncover more diverse and comprehensive vulnerabilities. Clients who utilize third-party services report **more effective flaw detection**, suggesting a broader emphasis on third-party involvement would enhance AI security and compliance.

3. **Expanding Focus Beyond Financial and Security Threats**: Although red teaming efforts naturally prioritize issues related to security and financial fraud due to their direct financial implications, it is essential to **extend the focus to areas that impact societal well-being.** This includes testing for vulnerabilities that could expose users to graphic violence, extremism, suicide/self-harm, and child safety concerns. Addressing harmful content is critical, as the damage from misinformation, deepfakes, and other malicious content extends beyond financial loss to significantly affect public trust and safety. Greater emphasis should be placed on these areas, aligning with recent NIST framework inclusions.

    - **Real life Case Dilemma:** OpenPolicy's partner, ActiveFence, raises a case study regarding the concerns that the red teaming of visual models for child safety, particularly regarding the generation of CSAM (Child Sexual Abuse Material). Due to the legal implications, organizations often avoid red teaming for visual CSAM, even though this approach could be effective in other abuse scenarios. **Discussions with regulators should explore potential exceptions that allow red teaming for CSAM in controlled, ethical ways to utilize this tool effectively against a critical risk area.**

4. **Up-to-date intelligence and Threat Actor Behavior:** Gathering and analyzing user feedback can help identify potential safety issues and areas for improvement, which is essential to identifying possible weaknesses and enhancing an organization's resilience. **Integrating intelligence** from up-to-date feedback, threat actor activities, and terminologies into safety processes would help better identify and mitigate distinctive risks.

Current intelligence indicates that bad actors use AI tools despite platform restrictions, employing techniques to bypass safeguards and generate harmful content, such as child pornography. Therefore, it is crucial to implement more robust methods to gather such intelligence and make it accessible to those responsible for safeguarding decisions. Emphasizing the need to monitor bad actor forums and behaviors should play a central role in red teaming. This approach ensures that red teaming efforts align closely with actual threat actor behavior rather than hypothetical scenarios.

5. **State-of-the-Art Red Teaming:** In line with the principles outlined in the **Streamlining Federal Cybersecurity Regulations Act**,[9] there is a need to adopt a holistic and unified safe-by-design approach to ensure that risks are monitored and mitigated at every stage of model development, release, and use. Beyond aligning and introducing a unified approach, continuous improvement and development are essential as the threat landscape constantly evolves and becomes more sophisticated. Therefore, ongoing investment in **understanding risks** and implementing appropriate safety solutions is crucial, including **expanding testing to cover more languages and abuse areas.** Collaboration between internal teams such as Trust & Safety, Responsible AI, and Security, along with independent evaluators and experts, is vital in fostering a robust and transparent AI testing framework.

   Statistical analysis is needed to define a sufficient volume and threshold for red teaming to ensure a model's safety. Providing organizations with guidelines on the appropriate extent and depth of red teaming would also be beneficial, as they enable companies to measure their efforts effectively and ensure robust defense mechanisms are in place.

   Moreover, a unified approach for AI testing should be considered, building on frameworks such as **MITRE-ATLAS,** which clearly distinguish between cybersecurity and AI-specific terms to support practical evaluation of AI systems. This includes developing and differentiating testing methodologies for security and "non-security harms," such as bias and discrimination, and establishing terminology for algorithmic flaws. Fostering **distinct definitions and testing methodologies for traditional cybersecurity threats versus AI algorithmic flaws** ensures a comprehensive evaluation of AI systems. Establishing clear standards and language for "AI

[9]See S.4630 - Streamlining Federal Cybersecurity Regulations Act
https://www.congress.gov/bill/118th-congress/senate-bill/4630/text

red-teaming" can further distinguish between adversarial and non-adversarial testing approaches, enhancing the efficacy and trustworthiness of AI technologies.

6. **Ongoing resilience methods: Ongoing red team exercises** should be an integral part of a comprehensive mitigation strategy to ensure resilience against evolving threats. Establishing a robust methodology with a pre-defined plan that outlines actions and methods, including the number of sessions, types of tests, and procedures for handling safe and unsafe responses, is essential. Additionally, implementing prompts with varying difficulty levels and conducting multiple iterations to observe variations in responses could better ensure the robustness of the AI systems and the organization's overall security posture.

7. **Incorporating Multi-Language and Regional Considerations:** It is crucial to recognize that red teaming limited to English-speaking contexts overlooks a substantial portion of the potential attack surface. Adversarial prompts in various languages are necessary, as models are often available in multiple languages, and harmful content may be more prevalent or varied in non-English languages due to a lack of focus. Therefore, AI red teaming should encompass diverse linguistic and regional perspectives to comprehensively address risks across different user groups and geographies.

/s/ Michelle Sahar
Michelle Sahar
Cybersecurity Policy Director, OpenPolicy