# Organisations adopt AI red-teaming to probe vulnerabilities and improve system safety



Organisations are increasingly adopting red-teaming practices to test and strengthen artificial intelligence (AI) models by intentionally probing their vulnerabilities and unwanted behaviours. This approach parallels traditional cybersecurity techniques, where ethical hackers simulate cyberattacks to identify and remedy weaknesses in network defences.

Red-teaming in AI involves rigorous testing to reveal the limits and potential risks of AI systems before they are widely deployed. Tori Westerhoff, principal director and AI red-teaming manager at Microsoft, described the process as “the tip of the spear” in uncovering vulnerabilities and “mechanised empathy” aimed at understanding and protecting users of high-risk generative AI technologies. She made these remarks during a webinar hosted by the Center for Security and Emerging Technology (CSET) in March.

At Microsoft, AI red-teaming focuses on evaluating how AI systems can behave, probing for potential soft spots and assessing how systems might fail or be exploited. This proactive testing helps identify problematic behaviours and enables developers to improve safeguards in AI deployments.

For non-profit organisations like the federally funded research and development centres managed by MITRE Corporation, AI red-teaming takes a slightly different form. Anna Raney, lead AI security engineer at MITRE, explained that their work often involves collaborating with government agencies to red-team AI systems that are recently acquired, near operational use, or already in operation. Their testing simulates real-world adversarial attacks on the entire AI-enabled system, encompassing the use case, stakeholders, and operational environment. During the same CSET webinar, Raney highlighted the comprehensive nature of MITRE’s approach to reveal vulnerabilities in operational AI systems.

The evolving nature of AI red-teaming has led to some ambiguity in understanding its scope and outcomes. Colin Shea-Blymyer, a CSET research fellow, noted that AI red-teaming is “an interactive and iterative process” to explore the robustness of AI systems, but the term can be “a little bit muddy.” He explained that red-teaming overlaps with capability elicitation — when testers try to uncover hidden functionalities not intended by the developers. For instance, if a model is claimed to be incapable of producing harmful outputs, a successful red-team effort might find ways to bypass safeguards and elicit such outputs, indicating system weaknesses. Conversely, it can reveal false positives where expected capabilities, such as performing mathematical proofs, fail.

A significant challenge in the field is the absence of standardised testing methodologies and reporting frameworks for AI red-teaming. Shea-Blymyer emphasised that differing approaches among red teams make it difficult to compare results across different AI products reliably. He suggested that while government-mandated standards may not be necessary, an authoritative organisation should advocate for unified reporting standards to clarify testing procedures and results industry-wide.

MITRE has contributed to addressing this gap through the creation of the Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS). This globally accessible knowledge base documents adversary tactics, techniques, and procedures targeting AI systems. Raney explained that ATLAS includes a section dedicated to case studies from both real-world incidents and red-teaming exercises. Such shared knowledge helps the community learn new tactics and improve overall AI security by expanding the collective understanding of AI adversarial threats.

Jessica Ji, a research analyst at CSET and moderator of the webinar panel, characterised AI red-teaming as being at an “awkward stage” of development. Many organisations are eager to adopt it, but the rapid evolution of AI means they are simultaneously developing the necessary tools, frameworks, and methodologies to conduct effective testing.

Westerhoff drew a parallel to the maturation of the cybersecurity field, which also faced early challenges in standardising practices and terminology. She stated, “We need strong perspectives and tools that are shared across the entire industry and are accessible to all so that we can start building our own dictionary … so that what we say about it means the same thing to everyone.” She cautioned that before codifying AI red-teaming into regulations, a clearer and more consistent industry understanding must be established to ensure effective and meaningful standards.

In summary, AI red-teaming is emerging as a critical method to validate and improve the reliability and safety of AI models by actively seeking out their flaws and capabilities. Through collaborative efforts, shared knowledge bases like ATLAS, and ongoing development of common frameworks, organisations aim to build trust in AI systems and mitigate risks associated with their deployment in security-sensitive environments.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://mindgard.ai/blog/what-is-ai-red-teaming> - Supports claims about AI red-teaming's role in identifying vulnerabilities through adversarial simulation and stress-testing AI systems, particularly in high-stakes environments.
2. <https://learn.microsoft.com/en-us/security/ai-red-team/> - Corroborates Microsoft's involvement in AI red-teaming practices and provides insights into industry-leading methodologies for securing AI systems.
3. <https://www.crowdstrike.com/en-us/services/ai-red-team-services/> - Validates the use of tailored adversarial scenarios in red-teaming, aligning with descriptions of operational AI system testing for real-world exploit prevention.
4. <https://www.dnv.com/article/ai-red-teaming-for-critical-infrastructure-industries/> - Supports the structured process of AI red-teaming (objectives, vulnerability testing, reporting) and mentions compliance-driven security enhancements.
5. <https://atlas.mitre.org/> - Directs to MITRE's ATLAS knowledge base (implied by context), which documents adversarial tactics and red-teaming case studies as described in the article.
6. <https://www.cset.georgetown.edu/> - Corroborates CSET's role in hosting relevant webinars and discussions on AI red-teaming challenges, including terminology standardization and maturity comparisons to cybersecurity.
7. <https://news.google.com/rss/articles/CBMivAFBVV95cUxNMU1HQXk3SXVoblVfVS14MG1EbDVBVWxxQy11TzdMWWtLRjZTWEZ6Q1luTjl0OFczT2lxQWdtc2tsOUJBdi1fR2t4QXlhREpZWDV2eVRCay1wUW9vQmhTeEVWOHpzZWtqTl8xZXRLYTJYSVJUMHRpczFSaHVUMV9GVElrRU1TUGRQZFhtTU9ZbVRZb3hGTlRKZUZJRmN2VXQzWVFZYmp0LU5OeVVJNjRZUkxXUW5abHg4RVh3UA?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data