AI Red Teaming in Practice: From Awareness to Action

Closing the Gap Between AI Adoption and Security Readiness

July 2, 2026

Dr. Alexander Zagnetko, KPMG Slovakia
Dr. Keri Pearlson, MIT Sloan School of Management
Leonidas Lykos, KPMG Austria
Dr. Griffin Reid, KPMG US
Billy Lawrence, KPMG UK
Kristy Hornland, KPMG US
Markus Hupfauer, KPMG Germany
Katie Boswell, KPMG US

©2026 Keri Pearlson. No part of this paper can be reproduced without reference to this copyright. Please note that AI was used in the process of producing this research report. GenAI tools were used to do some background research, which was then verified by our authors. GenAI tools were used to summarize and preliminarily analyze the data collected, and authors then reviewed and revised analysis for validation and clarity. GenAI tools were also used to polish some of the sections of this final paper, and the output reviewed and edited by our authors.

The authors wish to acknowledge support for early stages of this research from the Cybersecurity at MIT Sloan (CAMS) consortium at Massachusetts Institute of Technology (MIT). Sincere thanks to KPMG member firm professionals that contributed to this paper.

Reference for this paper: Zagnetko, A., Pearlson, K., Lykos, L., et.al. (2026). *AI Red Teaming in Practice: From Awareness to Action.* SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7133878

AI Red Teaming in Practice: From Awareness to Action

Abstract

AI introduces novel and amplified cybersecurity risks. Given the probabilistic nature of the systems, combined with the development process of an AI system, traditional security risks are combined with new vulnerabilities expanding the threat landscape. These new risks are not well-evaluated by traditional security testing tactics, techniques, and procedures (TTPs). Instead, developers and security professionals must leverage red teaming concepts, building AI red teaming capabilities and processes. Traditional red teaming has long served as a critical safeguard in cybersecurity, enabling organizations to identify and remediate weaknesses before adversaries exploit them. AI red teaming is the systematic practice of testing AI systems against these risks. Unlike traditional red team exercises, which are typically bound by technical exploits, AI red teaming expands the aperture to include behavioral vulnerabilities. While many organizations have well-established processes and budgets for traditional red teaming, AI red teaming remains an emerging capability. This paper reports on a study conducted to better understand how organizations use AI red teaming. We found that many of the organizations did not conduct AI red teaming, primarily because they were in the early stages of adopting AI which created different priorities. For those that were performing AI red teaming, the data showed that it was still in the early stages of adoption. Most organizations were selectively using these practices. The findings suggest a number of actionable insights for business, AI, Cyber and IT leaders and for those building or investing in AI systems.

Introduction: The Changing Risk and Threat Landscape

The increasing integration of artificial intelligence (AI) into business operations has introduced a new set of risks for organizations. These are risks that traditional security testing tactics, techniques, and procedures (TTPs) are not equipped to fully address. Instead, developers and security professionals must leverage red teaming and adapt traditional TTPs to fit the new AI environment.

AI red teaming is a discipline focused on systematically probing AI systems to uncover vulnerabilities in safety, robustness, privacy and security. Unlike conventional red teaming, which targets software, system, and human flaws, AI red teaming interrogates how intelligent systems behave under adversarial conditions, highlighting risks such as prompt injection, data poisoning, model inversion, and harmful outputs. Given the way AI systems in general, agentic and generative AI systems in particular, are growing in popularity, this discipline is critical to reducing new risks and vulnerabilities.

Although traditional red teaming and AI red teaming operate in different domains, they share a common philosophy: simulating adversarial conditions to proactively identify weaknesses. Both practices rely on the expertise of skilled assessors who adopt the mindset of attackers, and both ultimately aim to strengthen defenses before real adversaries can exploit vulnerabilities. However, while both practices share the goal of preemptively surfacing weaknesses, their domains diverge. Traditional red teaming strengthens systems, whereas AI red teaming not

only strengthens systems, but also ensures trust and resilience in decision-making processes driven by AI.

In this paper, we discuss the concept of red teaming, apply red teaming to the AI environment, and share observations and lessons learned from a survey of developers and c-level leaders for AI systems.

Traditional Red Teaming

Traditional red teaming has long served as a critical safeguard in cybersecurity, enabling organizations to identify and remediate weaknesses before adversaries exploit them. By simulating real-world attacks against networks, applications, infrastructure, and people, red team operators provide actionable insight into vulnerabilities, misconfigurations, and gaps in security controls as well as detection and response capabilities.

The value of traditional red teaming lies in its ability to provide organizations with a prioritized set of findings that drive remediation efforts, improve overall resilience, and ensure compliance with regulatory requirements. In this sense, red teaming has historically operated as a technical assurance mechanism, focused on safeguarding digital infrastructure against known classes of threats. Over time, it matured into a widely standardized and regulated practice, with consistent methodologies, specialized skillsets, and recognized importance at the enterprise level.

AI Red Teaming is Different

AI red teaming is the systematic practice of testing AI systems against these risks. Unlike traditional red team exercises, which are typically bound by technical exploits, AI red teaming expands the aperture to include behavioral vulnerabilities. These vulnerabilities may manifest in unintended outputs, systemic biases, or misuse potential; risks that could undermine trust in the system even when no conventional “bug” exists.

AI red teaming generally involves testing guardrails designed to prevent against harmful prompts (prompt injections / generating undesirable outputs / chaining adversarial prompts / hallucinations), evaluating content filters (bypassing content filters), probing for jailbreaks, and assessing resilience against adversarial examples, enhancing trustworthiness to ensure AI systems behave as expected or whether they can be manipulated to produce harmful, biased, or policy-violating outputs, even under adversarial pressure (Goldstein et al., 2023; OpenAI, 2023).

The results of AI red teaming provide organizations with insights not just into whether their AI systems are secure, but whether they are safe, resilient, and aligned with ethical and regulatory expectations, and highlight that AI systems can fail not because of broken code, but because of undesirable or unsafe behavior under adversarial conditions.

How Organizations Apply AI Red Teaming

Despite its growing importance, there is limited research on how organizations are currently approaching AI red teaming. Unlike traditional red teaming, which has become a well-established practice over decades, AI red teaming is just emerging and is in need of

standardized maturity models or industry benchmarks before it becomes common practice. There is a lifecycle to adopting AI red teaming practices, including early exploratory stages, developing structured methodologies, and establishing formal governance. In this study, we want to understand the progression through this lifecycle.

While many organizations have well-established processes and budgets for traditional red teaming, AI red teaming remains an emerging capability. It is our observation that enterprises are still exploring how to adapt traditional security assurance to AI systems, and few have standardized approaches. Current practices vary widely, from ad hoc internal testing by data scientists to structured engagements with emerging AI security firms.

A central challenge is that many organizations have not yet fully grasped how AI fundamentally alters the risk profile of their digital ecosystems. Traditional IT systems typically present deterministic vulnerabilities. Generative AI (GenAI) systems, in contrast, are, to some extent, indeterministic (sometimes called probabilistic), with outputs influenced by a vast number of factors that far exceed the spectrum of classical IT-systems. This creates scenarios where an adversarial action may succeed under certain conditions and fail under others. The behavior is better understood through the lens of chaos theory: the vast number of interacting parameters, model weights, architecture, training data nuances, prompt phrasing and context, tokenization quirks, hyper-parameter settings, context window sizes, and external data drift in RAG scenarios can amplify minute differences into qualitatively different outputs. In such a high-dimensional, semi-stochastic system, tiny, practically unobservable variations can determine whether an attack succeeds, making it effectively impossible to perform red team engagements with the same quality expectancy as for classical IT-systems. Consequently, AI red teaming must move from binary, annual tests to continuous, probabilistic assessment and adaptive mitigations that accept, embrace and manage a lot of uncertainty. The quantity and quality of AI red teaming directly influences the level of uncertainty but will most likely not remove it in the close future.

As companies race to deploy AI and embed critical decision-making processes, they must not only adopt new technical methodologies but also develop governance frameworks, threat models, and cross-functional teams to responsibly manage the novel risks of intelligent systems.

Defining AI Red Teaming

AI red teaming may mean different things to different stakeholders. For the purposes of this research, the following definition was developed to reflect the evolving nature and complexity of AI systems, and to distinguish this practice from traditional red teaming approaches:

> "*A specialized, proactive, and continuous practice where experts simulate adversarial attacks against AI ecosystems to discover unique vulnerabilities and weaknesses, and test response and detection capabilities, aiming to ensure the security, fairness, transparency, trust, and robustness of AI models and algorithms".*

While conventional red teaming typically involves simulating adversarial behavior against static systems, such as applications, networks, or physical infrastructure, the primary goal is to identify security gaps. In contrast, *AI red teaming* targets dynamic, learning-based systems that are

inherently probabilistic, adaptive, and often opaque. This shift in focus demands a broader lens that encompasses not only security, but also *fairness, transparency* and *trustworthiness*, which are essential pillars of AI assurance. Our definition of AI red teaming was designed to address this.

AI red teaming encompasses the entire AI ecosystem, including models, data pipelines, deployment environments and human-AI interaction. This comprehensive scope ensures that vulnerabilities are identified not only in code or infrastructure, but also in the behavior and decision-making of AI systems under adversarial conditions and/or misuse. The methodology is inherently iterative and continuous, reflecting the evolving nature of AI technologies and their exposure to new data, contexts and threat vectors. A sustained and proactive engagement is essential to building resilient and trustworthy AI systems, particularly as they become increasingly autonomous and embedded in critical decision-making across an organization.

AI red teaming is a critical practice for probing the safety, robustness, and ethical implications of AI in general, and genAI systems specifically. Originally rooted in military and cybersecurity domains, red teaming now plays a central role in AI governance, especially for large language models (LLMs) and multimodal systems.[1,2] The practice involves adversarial testing - where experts simulate attacks or misuse scenarios to uncover vulnerabilities. This includes prompt injection, data poisoning, model inversion, and more. Several leading AI and technology companies have institutionalized red teaming as part of their Responsible AI frameworks.[3]

Risks in AI Security

AI introduces novel and amplified cybersecurity risks. Given the probabilistic nature of the systems, combined with the development process of an AI system, which includes training the model and a feedback loop that continues to train the model, traditional security risks are combined with new vulnerabilities to create a different threat landscape. The Global Cybersecurity Outlook 2025 found that while 66% of organizations expect AI to impact cybersecurity, only 37% have processes to evaluate AI system security before deployment.[4]

As highlighted by the World Economic Forum[5,6] and Carnegie Endowment[7], some of the key risks include:

- **Adversarial attacks**: Manipulating inputs to deceive models (e.g., image perturbations, prompt injections).
- **Data poisoning**: Corrupting training data to embed vulnerabilities or biases.
- **Model inversion and extraction**: Reconstructing training data or replicating models from outputs.
- **Autonomy risks**: AI agents making decisions without human oversight, potentially leading to harmful actions.
- **Algorithmic manipulation**: Exploiting model behavior to achieve unintended outcomes (e.g., gaming recommendation systems).
- **Supply chain vulnerabilities**: Risks from third-party model components or datasets.

- **Lack of explainability**: Difficulty in auditing and understanding model decisions, which complicates risk mitigation.

To reduce risks from GenAI systems, organizations use AI Testing, Evaluation, Verification, and Validation (TEVV) frameworks. NIST's proposed standard[8] outlines a structured approach to TEVV, emphasizing:

- **Sector-specific testing**: Tailoring evaluation methods to the domain (e.g., healthcare, finance).
- **Lifecycle integration**: Embedding testing throughout development, deployment, and monitoring.
- **Diversity of methods**: From statistical robustness checks to adversarial testing and human-in-the-loop evaluations.

However, current models are often inadequate. Many TEVV methods are still experimental or lack consensus and therefore standards. Organizations use inconsistent approaches making benchmarking difficult and fragmented. Some models focus only on performance metrics and ignore ethical or security dimensions making them limited in scope. Benchmarks need to evolve as AI model capabilities evolve. Finally, testing processes are primarily manual, lacking automation, which get in the way of scalability and responsiveness due to:

- **Immaturity of standards:** Many TEVV methods are still experimental or lack consensus.
- **Fragmentation:** Organizations use inconsistent approaches, making benchmarking difficult.
- **Limited scope:** Some models focus only on performance metrics, ignoring ethical or security dimensions.
- **Lack of automation:** Manual testing processes hinder scalability and responsiveness.

MITRE's AI Maturity Model[9] offers a complementary framework for assessing organizational readiness and capability in AI adoption. It includes six pillars (e.g., ethical use, data, performance) and five maturity levels, helping organizations align testing practices with strategic goals.

AI Red Teaming is an Important Component of AI security

Historically, red teaming has involved simulating cyberattacks to test the resilience of IT systems, networks, and organizational defenses. It focuses on known vulnerabilities, exploits, and adversarial tactics within traditional digital infrastructures. AI red teaming targets the unique characteristics and vulnerabilities of AI systems. According to MITRE[10], AI red teaming involves:

- **Adversarial testing** of AI models, including probing for bias, robustness, and susceptibility to manipulation (e.g., adversarial examples, prompt injection).
- **Simulating misuse scenarios**, such as model inversion, data poisoning, or unauthorized model replication.
- **Evaluating emergent behaviors** in complex systems, especially in generative models and autonomous agents.

- **Assessing sociotechnical risks**, including misinformation propagation, ethical violations, and unintended consequences.

For example, MITRE's ATLAS framework[11] extends the ATT&CK model to AI systems, cataloguing real-world adversarial tactics and techniques specific to AI. This includes threat emulation tools like CALDERA plugins tailored for AI-enabled systems.

To summarize, the key differences between "traditional" and AI red teaming are as follows:

- **Scope**: AI red teaming includes both technical and sociotechnical dimensions.
- **Attack surface**: AI systems have broader and less predictable attack surfaces.
- **Tools and frameworks**: AI red teaming uses specialized tools not found in traditional red teaming.
- **Focus on model behavior**: Unlike traditional red teaming, which targets infrastructure, AI red teaming scrutinizes model outputs and decision-making.

Red teaming is increasingly recognized as a sociotechnical challenge. Gillespie et al. argue that red teaming must account for values, labor dynamics, and psychological impacts on testers.[12] The human element - biases, blind spots, and emotional toll - is central to understanding the effectiveness and ethics of red teaming.[13] Microsoft's internal study of red teaming over 100 generative AI products offers eight key lessons:[14]

- Red teaming is not equivalent to safety benchmarking.
- Automation can expand coverage but cannot replace human insight.
- LLMs amplify existing risks and introduce novel ones.
- Responsible AI harms are pervasive but hard to quantify.

These insights underscore the need for continuous, adaptive red teaming strategies. Majumdar et al. propose a dual-level framework: macro-level system red teaming and micro-level model red teaming.[15] This approach emphasizes the need to examine emergent risks and systemic vulnerabilities beyond isolated model flaws. Other works, like the end-to-end overview by Alberto Purpura et al., detail the components of a red teaming system, including threat modelling, scenario generation, and mitigation pipelines.[16]

The largest AI red teaming exercise to date occurred at DEF CON, involving thousands of participants testing LLMs for harmful outputs.[17] While such events demonstrate scalability, they also reveal challenges in standardizing methodologies and interpreting results. Bajema (IEEE Spectrum) notes that red teaming, while useful, cannot substitute for structural regulation.[18] The pace of AI development often outstrips the ability of red teams to identify and mitigate risks in real time.

Certain psychological effects of AI red teaming are underexplored. Zhang et al. call for interdisciplinary research into labor conditions, fairness, and mental health impacts.[19] This aligns with broader concerns in HCI and CSCW literature about the ethics of content moderation and algorithmic auditing.

Emerging Trends and Challenges in AI Red Teaming

Several trends are shaping the future of AI red teaming. Especially relevant to our research are the trends of tooling and automation where new software packages aim to streamline AI red teaming workflows[20], multimodal AI red teaming where models like OpenAI's Sora expand into video, red teaming must evolve to address deepfakes and synthetic media[21], policy integration with governance where regulation and appropriate use of AI red teaming continue to be debated[22], and the extension of AI red teaming to agents.

Tooling and Automation: A clear trend is the move toward automation in AI red teaming. Traditional red teaming was largely a manual, human-driven effort – experts brainstorming attacks and trying them one by one on the target system. But with AI models (especially LLMs), the space of possible inputs and behaviors is enormous, and manual testing cannot cover it comprehensively. Several publications have noted this scalability challenge and the opportunity to leverage AI itself to test AI. Shruti Srivastava et al. (2026) provide a Systematic Review of Algorithmic Red teaming[23] observing that "manual red teaming is resource-intensive, slow, and cannot keep pace with evolving threats," which is driving adoption of automated red teaming. Automating red teaming involves using scripts or even AI models to generate adversarial inputs, systematically explore edge cases, and even evaluate model outputs for signs of failure. This doesn't replace human ingenuity, but augments human red teams by handling the brute-force exploration and allowing experts to focus on high-level strategies.

Multi-Model AI Red Teaming: Another trend is the broadening of what "AI red teaming" encompasses – from focusing on a single model's behavior to examining whole AI-powered systems, especially those exhibiting agentic behavior. In early generative AI deployments (circa 2022–23), red teaming was often scoped to one model at a time, largely testing, say, a chatbot's prompt-response behavior in isolation. Now, as AI is integrated into complex products and allowed to take actions, red teams are simulating attacks that target the full sequence of model decisions and tool uses.

Integrating Policy with Governance: A third trend is the institutionalization of AI red teaming. AI red teaming is being explicitly built into governance frameworks and expected by policymakers. This is clearly illustrated by the EU's Artificial Intelligence Act. Article 9 of the EU AI Act (which applies to "high-risk AI systems") mandates that providers must perform a risk management process that includes identifying and mitigating risks through testing, including with adversarial techniques. The penalties are severe – up to 6–7% of global annual turnover for non-compliance. So regulatory pressure is explicitly driving companies to adopt red teaming. This was anticipated in various policy discussions last year, and in early 2026 it became a reality as the AI Act moved into enforcement.

Extension to AI Agents: Another major trend is the extension of AI red teaming to autonomous AI agents. By late 2025, systems like LLM-based agents (which can execute code, use tools, or interact with other agents) have grown popular in experimental applications – and they bring new risks not seen in single models. Academic and professional publications have identified threats unique to these agents, including tool misuse, self-modifying behaviours, leaking or corrupting long-term memory, and coordinated multi-agent misbehaviour. In response, the first frameworks for "agentic AI" red teaming appeared.

Google's 2026 progress report places heavy emphasis on how their AI red-teaming programs have expanded and how they feed into safer deployments. Interestingly, Google also launched an AI Vulnerability Reward Program (VRP) in 2025, effectively a bug bounty for AI issues like jailbreaking and data leaks. This moves AI red teaming partly into the public domain by rewarding external researchers who find vulnerabilities, analogous to how software security is handled. It's another sign of institutionalization: treating AI issues like security bugs to be reported and fixed.[24]

Despite growing adoption, challenges remain for using and managing AI red teaming. Successful AI red teaming is as much about process and governance as technology. Recent peer-reviewed research portrays AI red teaming as an increasingly mature, multi-faceted discipline at the intersection of AI technology, security engineering, and ethical governance. Far from a mere technical exercise, red teaming is being reimagined as a holistic process – one that engages diverse human expertise and automated tools in tandem, that spans the entire AI system lifecycle, and that feeds into continuous risk mitigation strategies. This evolution has been driven largely by the urgent need to tame the tendency of generative AI models to produce unintended, harmful, or exploitable outputs.

Key organizational practices and recommendations include:

- Embedding red teaming in the AI development lifecycle.[25]
- Create multidisciplinary and diverse teams: There is a clear consensus that red teams should include a diverse range of expertise.[26, 27]
- Ensure AI red teaming leads to actual improvements. It's not enough just to find errors and risks, but organizations must fix them as part of the red teaming processes. Feedback loops are a primary component of successful red teaming.[28]
- Build in continuous learning and adaptation: AI red teaming is part of a broader shift toward continuous AI risk management.[29]

Methodology

Based on the background reading and questions that arose from our survey of previously published articles and research on AI red teaming, the research question driving this study was:

> ***How are organizations leveraging AI red teaming in their AI security programs, targeting CISO and CIO levels?***

To explore this question, we generated a number of hypotheses, listed below:

1. *Relatively few companies are performing or are preparing to perform AI red teaming*
2. *AI red teaming is not currently a topic of discussion for C-level executives*
3. *AI security in general is not seen as an integral component for the overall success of an AI strategy*
4. *Companies performing AI red teaming:*
    a. *Do it as part of their secure-by-design processes and protocols*
    b. *Have built this capability into their regular quality control/security check processes*

    c. *Do not assess the maturity of their AI red teaming procedures regularly, thus they are most likely incomplete.*
    d. *Have red teams that work regularly with the AI Development teams.*
    e. *Have a single point of contact responsible for overseeing it.*
    f. *Regularly report results to the AI Development teams, the CIO and other C-level executives (and the Board?).*
5. *Companies not doing AI red teaming:*
    a. *Have a different way to test their AI systems for vulnerabilities*
    d. *Are not aware of benefits / other risk-domains covered by red teaming as compared to penetration testing.*

To test the research hypotheses, a survey questionnaire was created using the Qualtrics online platform to collect data and to build a theory. The survey comprised 22 questions in total, including demographic questions (e.g. role, country, industry, and organization size). Appendix A has a copy of the survey questions.

As discussed in the introduction, 'AI red teaming' is a relatively new domain and term. To this end, the definition was provided in the survey to ensure all participants were adequately informed of its meaning in the context of this research project. Furthermore, survey participants were informed that all responses would remain anonymous. They were given the option of sharing their contact details with the research team to receive the findings.

The survey was conducted from November 2025 through the end of April 2026. It is important to note that responses were collected at irregular intervals throughout this period; the research was presented at a series of focus-groups on AI security, primarily held in the US. Most responses were gathered during or after these focused sessions, when participants were provided with the survey link and QR code and invited to contribute to this research and share their insights. Accordingly, all but one of the survey respondents were based in the US, underscoring the geographical boundaries and limitations of this study. In the Discussion section, the limitations of this research are examined in greater detail, noting the extended data collection period, the evolving contextual environment during that time, and the relatively small sample size. Consequently, the findings are considered valuable for generating hypotheses and guiding future research, but probably not generalizable to a larger, more heterogeneous sample.

We received a total of 99 survey responses. Of these, 34 were deemed sufficiently complete and meaningful for analysis, defined as responses that addressed the survey questions and did not consist of blank, incomplete, nor unintelligible content. Due to the sensitive nature of AI red teaming, we anticipated that some potential respondents might be hesitant to participate fully. To encourage participation and ensure privacy, all demographic questions were made optional. Consequently, the demographic analysis is based on the 29 respondents who chose to provide information about their professional role and organization.

In terms of industry representation, respondents were predominantly drawn from Higher Education, Healthcare/Biotechnology, Financial Services, and Professional Services/Consulting, as shown in Figure 1 below.

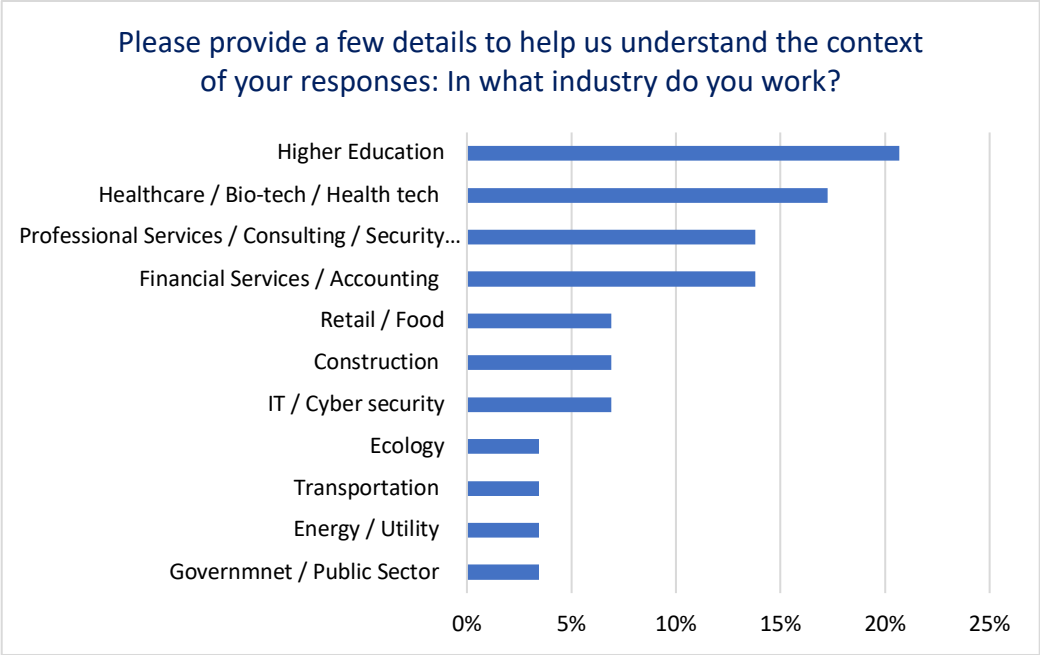


Figure 1 – Demographic question included in survey questionnaire

Approximately 86% of the respondents work in IT/Cybersecurity and/or broader Technology roles (see Figure 2 below). A notable subset of these respondents (approximately 45%) holds managerial or executive-level positions within their organization, such as Head of IT/Security, CIO, CISO and similar roles. This profile is consistent with that of the subject matter experts who participated in the focus-groups.

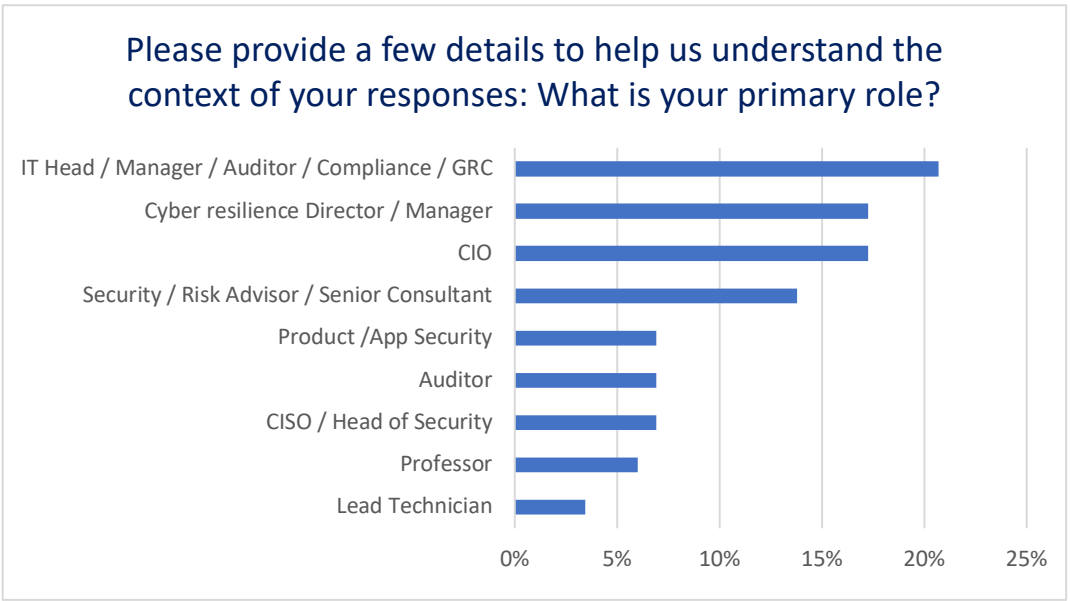


Figure 2 – Demographic question included in survey questionnaire

Considering that some organizations may still find themselves in the early stages of their AI adoption journey, this research was also interested in exploring why organizations may not be conducting AI red teaming. This logic was applied to the survey structure and question sequence; the respondent was prompted to explain the reasons for this in their organization.

Results

The research yielded a series of noteworthy findings in the field of AI red teaming
In this section, we summarize the results of our data analysis. The data collection began with a series of questions about AI use and governance. When asked about general enterprise AI solutions, 53% of the respondents said they were exploring or piloting enterprise AI solutions and 32% are actively using AI enterprise solutions. In another question, we asked how many AI systems or models were currently in use, and we learned that most of the respondents (68%) have at least 1 AI system or model actively in use.

AI Governance Practices
Our survey asked a number of questions about the respondent's governance practices. Specifically, we asked if there was a dedicated policy or governance framework. Over half (56%) of the respondents shared that they have a formal and enforced policy or are currently developing a policy. Almost 22% shared that they rely on general IT governance to cover their AI use.

When asked about unauthorized AI tools and shadow AI, only 21% said they maintain a strict blacklist and/or monitor use of unauthorized AI tools or shadow AI. Informal controls are in place for about 1/3 of the respondents (30%) and a further 21% of respondents shared that they do not manage unauthorized AI tools or apps use at all. Another 18% responded that they are planning to implement controls at the present time.

Next was a question about how AI systems were developed or acquired. Respondents were asked to select all that apply. The majority of responses for this question (58%) shared that they use external AI solutions or systems. Twenty-seven percent shared that they developed AI solutions internally, and 15% have used outsourcers to provide AI systems. It was interesting that 15% responded that their organization don't use AI systems at all at the present time.

Machine learning operations (MLOps) was also queried. By far the most frequent response was that they did not have MLOps practices in place (36%). Only 15% have fully implemented MLOps practices and another 12% have partially implemented MLOps.

AI Red teaming Practices
As illustrated in Figure 3 below, many of the respondents (26%) did not conduct AI red teaming at the time of this study. Another 24% did not know if their organization conducted AI red teaming. For those not conducting AI red teaming, the reasons why varied. Many cited limited resources (63%). Another 45% said they relied on third party providers, presumably the vendor who is providing the AI solution themselves. At the same time, 63% of respondents said they were at an early stage of AI adoption and had not yet implemented AI red teaming to test solutions. Interestingly, a minority of respondents, 18%, said they only focused on regulatory compliance.

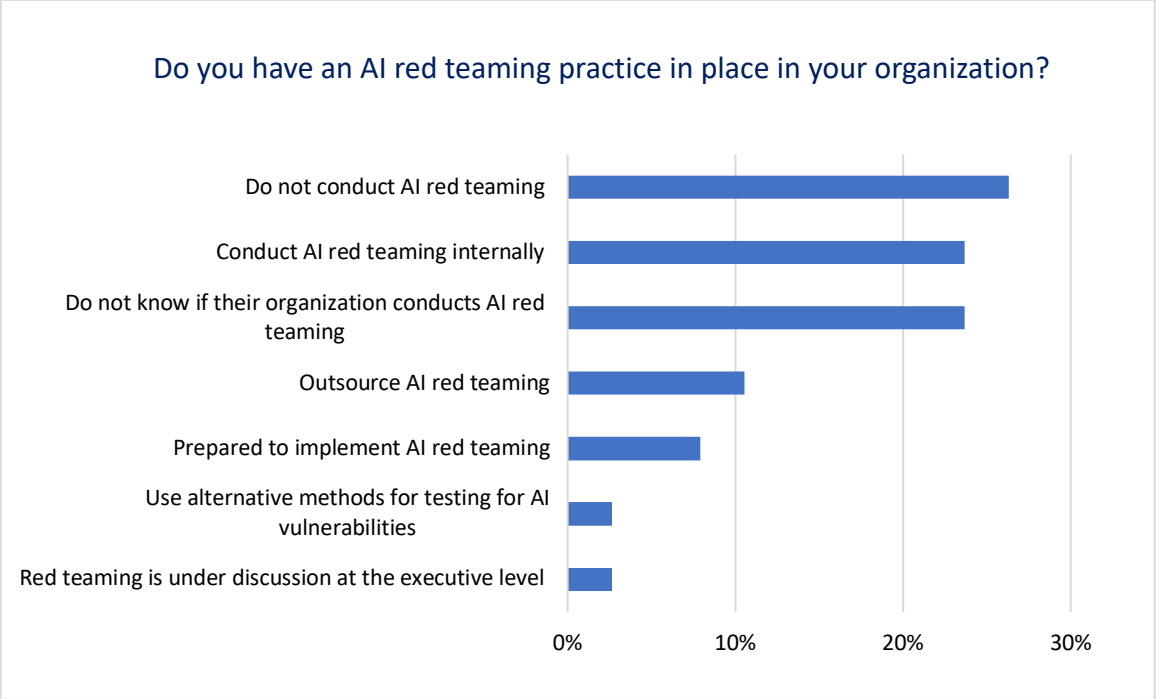


Figure 3 – Question #13 in survey questionnaire (see Appendix for full survey)

For those that do conduct AI red teaming, we asked a number of probing questions to better understand their current practices. When asked what percent of their AI systems undergo AI red teaming today, 40% said that less than 25% of their organization's AI systems undergo this level of security check. Sixty percent responded that more than 25% of their AI systems undergo AI red teaming today (with 30% responding that more than 75% of their systems are tested this way).

When asked what types of vulnerabilities or attacks are tested during AI red teaming exercises, the respondents were given a list of choices. A significant majority (88%) percent selected model/data extraction as the most common type tested. Inference attacks (63%), data poisoning (50%) and evasion attacks (38%) were also identified.

AI red teaming exercises were conducted annually by 56% of the respondents and monthly or more frequently by 33% of the respondents. One respondent said that AI red teaming exercises were done on an ad-hoc basis.

NIST AI RMF and OWASP AI Guidelines[30] were the most frequently cited (54%) framework or guidelines used for AI red teaming. MITRE ATLAS was mentioned by 36% of the respondents and 27% said they also used an internal framework.

AI Red teaming Management

Respondents who conduct AI red teaming were also asked about the integration with other development and security processes. Forty-two percent said that AI red teaming was part of their regular quality and security checks, while the same number, 42%, said that AI red teaming

was conducted occasionally or experimentally. A small minority, 17%, said that they had fully integrated AI red teaming into their secure-by-design protocols.

We wanted to understand how frequently AI red teaming and AI development teams collaborated. Half (50%) of respondents said that coordination was just an occasional occurrence. Another 25% said that the teams work closely and regularly together. Just 17% said the teams work independently.

When asked who was responsible for overseeing AI red teaming, 60% said that there was a dedicated point of contact or team responsible. For 30% of the respondents, it was a shared responsibility across departments.

Finally, we asked who received the results of AI red teaming exercises, and allowed respondents to select all that applied. The majority of the respondents, 58%, said the CISO receives the results. Risk management or internal audit was identified as recipients by 42% of respondents, and 33% of respondents said the CIO received the results.

Discussion

There is a notable discrepancy between the rapid uptake of AI technologies and the maturity of supporting governance and lifecycle controls. Most enterprises are actively adopting or experimenting with AI, yet relatively few have fully instituted comprehensive oversight and management processes such as robust AI governance frameworks or mature MLOps practices. This gap suggests that organizations remain in an early maturity phase: the pace of innovation has outstripped the evolution of formal controls. As a result, many firms face challenges in ensuring responsible AI operation at scale, indicating potential vulnerabilities in risk management and compliance.

A related issue is limited visibility into “shadow AI” activities – the use of ungoverned or unsanctioned AI tools. The data indicates that while some companies attempt to handle unauthorized AI usage through informal measures, a significant segment either lacks controls or is only beginning to plan them. This uneven oversight of shadow AI implies that many organizations may not fully know what AI applications are being used across the enterprise, which could introduce unmitigated risks and compliance issues. The need to better monitor and manage unsanctioned AI tools emerges as a clear priority to close this governance gap.

Adoption of AI red teaming remains nascent and inconsistent. Only a minority of surveyed organizations have established internal AI red teaming programs, with many others either not conducting these exercises or being unsure if any occur. This limited uptake – driven more by practical constraints like early-stage AI initiatives and resource limitations than by a lack of belief in red teaming’s value – underscores that adversarial testing for AI security is still far from a mainstream practice. In effect, comprehensive threat testing is often deferred or outsourced, reflecting an ad hoc rather than institutionalized approach.

Among the few organizations that do conduct AI red teaming, the scope and frequency of testing vary widely, further evidencing a lack of standardization. Some red teams focus their efforts on only a small subset of critical AI systems, whereas a few manage to cover a broad

swath of their AI portfolio. Likewise, the cadence is typically infrequent: annual testing is the most common practice, with only isolated cases of more frequent (e.g. monthly) exercises. These patterns imply that AI red teaming is often regarded as a specialized, periodic assessment rather than a continuous, secure-by-design process integrated into the AI development lifecycle. Only rare instances show AI red teaming embedded as a routine design-time safeguard, indicating an opportunity for greater integration into standard engineering and security workflows.

The feedback on existing AI red teaming efforts reveals enterprise priorities: tests predominantly target risks of model or data compromise – specifically preventing sensitive information leaks or model intellectual property extraction – whereas challenges like adversarial evasion receive comparatively little attention. This emphasis on data extraction and inference vulnerabilities suggests that organizations perceive data confidentiality and model misuse as the most immediate threats to address, possibly because these scenarios are more tangible or have clearer business impact. In contrast, fewer respondents are proactively probing evasion or adversarial manipulation attacks, which could indicate either a lower perceived likelihood of these sophisticated threats or current limitations in expertise and tools to test them.

As the AI Bill of Materials (AI BOM) becomes an increasingly common prerequisite for the implementation of new AI solutions, AI testing and red teaming are also likely to become important components of business continuity and risk mitigation strategies. Organizations may increasingly seek confirmation that new AI modules or solutions have undergone appropriate testing and attack simulations, supported by documented results, before they are approved for deployment.

AI Amplifies the Existing Attacker–Defender Asymmetry

The market is evolving rapidly. Recent advancements in AI are blurring the line between offensive and defensive applications. As the field of AI red teaming progresses, new tools are increasingly emerging as double-edged swords—expanding what is possible for both malicious actors and responsible defenders of AI systems

AI is reshaping cybersecurity by increasing the offensive leverage available to attackers faster than most organizations can strengthen their defensive capacity. The core asymmetry remains unchanged: attackers only need to identify one viable path to compromise, while defenders must continuously protect a complex, dynamic, and interconnected digital environment. AI intensifies this imbalance by enabling greater speed, scale, persistence, and variation across the attack lifecycle.

Attackers can use AI-enabled capabilities to continuously monitor public-facing assets, software releases, exposed APIs, leaked credentials, misconfigurations, and third-party dependencies. They can also generate and test multiple attack variations, correlate weaknesses across cloud, identity, application, infrastructure, and supply chain layers, and tailor phishing, impersonation, or social engineering attempts to specific targets. In this context, offensive AI can be

probabilistic and opportunistic: many paths can be tested, and failed attempts can simply be abandoned. Defensive AI, by contrast, must be accurate, contextual, explainable, governed, and operationally reliable.

This creates a fundamental challenge for enterprise security teams. Modern organizations operate across cloud workloads, SaaS platforms, APIs, microservices, CI/CD pipelines, identity providers, data platforms, legacy systems, third-party integrations, human workflows, non-human identities, and increasingly AI agents. For attackers, this complexity creates opportunity. For defenders, it creates noise. Many harmful activities may initially resemble legitimate behavior: a compromised account may use valid credentials; a malicious API call may look like normal system integration; and an AI agent may use approved tools and permissions while pursuing an unintended or harmful outcome.

Agentic AI sharpens this asymmetry further. Unlike traditional generative AI, which primarily supports content generation, code assistance, phishing, or malware-related tasks, agentic AI can plan, call tools, execute workflows, interact with systems, and operate with a degree of autonomy. As a result, attackers may increasingly use AI not only for advice, but to automate parts of reconnaissance, vulnerability discovery, exploitation, and social engineering. At the same time, organizations may unintentionally introduce new attack surfaces by deploying internal agents with access to systems, data, APIs, tools, and business processes. This means agentic AI does not simply increase the number of vulnerabilities. It increases the number of autonomous or semi-autonomous pathways through which vulnerabilities, permissions, tools, identities, and business logic can be combined. Risks such as goal hijacking, tool misuse, privilege abuse, insecure inter-agent communication, memory or context poisoning, cascading failures, and rogue agent behavior illustrate that agentic AI is not just another application layer. It can become an orchestration layer across the enterprise environment.

Defensive use of AI is therefore inherently more constrained than offensive use. Defenders cannot block everything suspicious without disrupting business continuity. They cannot tolerate excessive false positives, because security teams are already overloaded. They cannot allow automated response to act too aggressively, because it may interrupt critical processes. Defensive AI must operate within governance, privacy, auditability, regulatory, and operational boundaries. It must determine whether behavior is malicious or merely unusual, whether an API call is expected, whether an agent is acting outside its intended scope, and whether a vulnerability is genuinely exploitable in a specific environment.

A related concern is that AI-driven vulnerability discovery may create more findings than organizations can realistically validate, prioritize, and remediate. The limiting factor may shift from identifying weaknesses to determining which findings are real, exploitable, material, and urgent. More vulnerability discovery does not automatically translate into better security if remediation capacity, responsible disclosure processes, patch management, and contextual risk assessment cannot keep pace. The risk is that AI industrializes vulnerability discovery faster than organizations can industrialize remediation.

AI red teaming reflects the same dual-use paradox. It is essential for testing models, agents, workflows, safeguards, and controls, but the methods used to identify weaknesses can also be misused to bypass protections, manipulate agents, or identify vulnerable configurations. The

answer is not to avoid AI red teaming, but to govern it carefully through clear authorization, controlled access, logging, monitoring, human oversight, responsible disclosure, separation from production environments, limits on autonomous execution, and strong identity and permission controls.

AI red teaming has transitioned from an experimental practice to a necessary component of deploying AI responsibly in business settings in 2026. We have seen AI red teaming become expected parts of the due-diligence processes and significant advancements in how to execute AI red teaming with new tools and collaborative approaches.[31,32]

Limitations of This Study

As we share and interpret the findings from this study, it's important to note the limitations of this research. Given the relatively small sample size of respondents and the fact that most of them are from the US, we note that our findings are useful for generating hypotheses for future study, but probably not generalizable to a larger, more heterogeneous sample.

Further, the study was conducted over a number of months which also saw much change in the AI field. The state of AI has steadily changed over the term of our study. There have been many advances in use cases, deployment, and production of GenAI. There has been significant growth in interest and popularity of agentic AI. And most organizations have increased their maturity and experience with AI development and practices. These point to a limitation of this study: interpreting our findings must be viewed within the context of the dynamic backdrop occurring during our data collection and analysis phases. It would be inappropriate to use our findings as a definitive description of the current state of AI security. Rather, our findings are useful for hypotheses generation and as indicators of areas for managerial action and future research.

Overall, AI does not make cyber defense impossible, but it raises the minimum maturity required to defend effectively. Organizations with weak asset management, identity governance, logging, segmentation, vulnerability management, incident response, or third-party risk controls may find that AI amplifies their exposure more than their resilience. To benefit from AI defensively, organizations need strong cyber foundations, contextual visibility, disciplined governance, and the ability to prioritize and act on findings at operational speed.

Actionable Insights

Define enterprise priorities and objectives for AI red teaming.
What does the organization expect AI red teaming to achieve? Priorities may include assessing risks related to model manipulation, data compromise, prompt injection, adversarial evasion, sensitive information leakage, misuse of AI agents, or broader system-level resilience. Clearly defined objectives will help ensure that red-teaming activities are targeted, measurable, and aligned with enterprise risk appetite.

Embed AI red teaming into AI acceptance and approval processes.
AI red teaming forms part of the criteria used to assess whether AI systems are suitable for deployment. Although the benefits of such testing may not always be immediately visible, the potential risks associated with untested AI vulnerabilities are significant. Red teaming can help identify model-specific weaknesses, misuse scenarios, data exposure risks, and unintended system behaviors that traditional security assessments may not capture.

Treat AI red teaming as a continuous secure-by-design practice.
Organizations must integrate AI red teaming throughout the AI system lifecycle, rather than applying it as a periodic or standalone security exercise. This includes incorporating adversarial testing during design, development, validation, deployment, monitoring, and major model or system updates.

Formalize AI red teaming within enterprise AI policies and operating procedures.
Mandate AI red teaming in relevant AI governance, cybersecurity, model risk management, and secure development policies. It must be treated as a standard operating procedure for relevant AI use cases, rather than an ad hoc or optional security activity.

Establish clear ownership for AI red teaming capabilities.
Invest in AI red teaming ownership early to find ways to incorporate it into all relevant steps in the security life cycle. This may involve creating an internal role or function responsible for developing internal expertise, coordinating testing activities, engaging external vendors, and integrating red teaming into the broader security and AI lifecycle. Without clear ownership, AI red teaming is unlikely to receive appropriate prioritization, resourcing, or executive attention.
Build governance and red teaming processes to include shadow AI.
Unsanctioned or informally adopted AI tools are present in most organizational environments. AI governance, risk management, and red-teaming processes must therefore be designed to account not only for formally approved AI systems, but also for shadow AI usage across the enterprise.

Adopt an adversarial mindset when designing AI red teaming processes.
AI red teaming must be based on realistic threat models and adversary behaviors. Organizations can ask what malicious actors would seek to exploit, how AI systems could be manipulated, and how attacks might affect business operations, data integrity, decision-making, or customer trust. If organizations do not proactively test their AI systems, adversaries may effectively perform this testing for them — but for malicious purposes. This mindset will guide AI red teaming activities towards more realistic threats and vulnerabilities.

Develop an internal AI red teaming framework while monitoring emerging standards.
Standardized AI red teaming methodologies, maturity models, and industry benchmarks remain relatively early in their development. Monitoring emerging regulatory, industry, and academic efforts, while selectively incorporating the most relevant practices into their own internal AI red teaming frameworks, allows organizations to build practical capability now, while remaining adaptable as the field matures.

Conclusion

Traditional red teaming remains a critical and well-established practice for assessing the security of conventional IT systems, which are largely governed by deterministic logic despite being data-dependent. However, AI red teaming addresses an emerging gap by focusing on the distinct risk profile of intelligent, probabilistic models, whose behavior is shaped by data-driven learning and may exhibit non-deterministic and unintended outcomes. For organizations, the challenge lies not only in understanding these differences but in preparing for them including developing new methodologies, building cross-disciplinary teams, and focusing energy on governance structures that ensure trust in AI systems.

Ultimately, just as traditional red teaming became an organizational norm, AI red teaming must become an operational and strategic imperative for enterprises deploying AI at scale.

The evolution from traditional red teaming to AI red teaming represents more than a shift in technical practice; it signals a transformation in how organizations must conceptualize and manage security, risk, safety, and trust. While traditional red teaming hardens infrastructure against known threats, AI red teaming extends assurance into the realm of behavior, alignment, and trustworthiness. For organizations, the challenge is not only technical but cultural and strategic: ensuring leadership support, funding, visibility, and cross-functional skillsets to operationalize AI red teaming at scale

There is an urgent need for empirical research that clarifies where organizations stand in this lifecycle, what obstacles impede progress, and which strategies most effectively accelerate maturity. Just as traditional red teaming matured into a normalized component of enterprise security programs, AI red teaming must evolve into a structured and optimized practice if organizations are to responsibly and securely deploy intelligent systems.

Appendix: The questions used in our survey instrument

Demographic questions – Please provide a few details to help us understand the context of your responses:

i. In what industry do you work?
ii. In what country do you work?
iii. What is your primary role?
iv. Is your company a small / medium ($50M – $1B) or large (>$1B) company?
v. Do you belong to the OWASP community?
vi. Which other AI security standards or groups do you participate in?

1. What is your organization's current use of enterprise AI solutions?
    - We actively use enterprise AI solutions
    - We are exploring or piloting AI solutions
    - We do not use AI solutions
    - I don’t know

2. Do you have a dedicated AI policy or governance framework?
    - Yes, we have a formal and enforced policy/framework
    - Yes, we have an AI policy/formal framework, but it is not enforced
    - We are developing one
    - No, but we follow general IT governance
    - No policy in place
    - I don’t know

3. How does your organization manage unauthorized AI tools/shadow IT?
    - We maintain a strict blacklist, formal controls, and monitoring system
    - We have informal controls in place
    - We are planning to implement controls
    - We do not manage this
    - I don’t know

4. How are AI systems developed or acquired in your organization? *(Select all that apply)*
    - Developed internally
    - Adapted from external solutions
    - Outsourced
    - We do not use AI systems
    - I don’t know

5. Do you have MLOps (Machine Learning Operations) practices in place?
    - Yes, fully implemented
    - Partially implemented
    - Planning to implement
    - No
    - I don’t know

6. How many AI systems/models are currently in use in your organization?
    - None
    - 1–5
    - 6–10

- More than 10
- I don't know

7. AI red teaming is a specialized, proactive, and continuous practice where experts simulate adversarial attacks against AI ecosystems to discover unique vulnerabilities and weaknesses, aiming to ensure the security, fairness, transparency, trust, and robustness of AI models and algorithms. Do you have such practice in place? *(Select all that apply)*
   - We conduct red teaming internally
   - We outsource red teaming
   - We are preparing to implement red teaming
   - It is under discussion at the executive level
   - We use alternative methods for testing AI vulnerabilities
   - We do not conduct red teaming
   - I don't know

8. Since you responded that you do not conduct AI red teaming, what are the main reasons? *(Select all that apply)*
   - Lack of awareness or understanding
   - Limited resources or expertise
   - Early-stage AI adoption
   - Rely on third-party providers
   - No clear ROI or business case
   - Focused only on regulatory compliance
   - Concerned about uncovering unmanageable risks
   - I don't know

9. What percentage of your AI systems undergo red teaming assessments?
   - 0%
   - <25%
   - 26–50%
   - 51–75%
   - > 75%
   - I don't know

10. What types of vulnerabilities or attacks are tested during red teaming? *(Select all that apply)*
    - Data poisoning
    - Evasion attacks
    - Inference attacks
    - Model/data extraction
    - Other
    - None
    - I don't know

11. (*If the last two answers above do not apply*) How often are red teaming exercises conducted?
    - Monthly or more frequently
    - Quarterly
    - Annually
      - On an ad hoc basis
      - Never

- I don’t know

12. Which frameworks or guidelines do you follow (or plan to follow) for AI red teaming or security? *(Select all that apply)*
    - NIST AI RMF
    - OWASP AI Guidelines
    - NVIDIA AI Red Team Framework
    - MITRE ATLAS™
    - Internal framework
    - Other industry frameworks
    - None
    - I don’t know

13. How integrated is/will be AI red teaming in your development and security processes?
    - Fully integrated into secure-by-design protocols
    - Part of regular quality/security checks
    - Conducted occasionally or experimentally
    - Not integrated but under consideration
    - Not integrated and not planned

14. How do your red teaming and AI development teams collaborate?
    - They work closely and regularly
    - They coordinate occasionally
    - They operate independently
    - No red teaming team exists
    - I don’t know

15. Who is/will be responsible for overseeing AI red teaming in your organization?
    - A dedicated point of contact or team
    - Shared responsibility across departments
    - No clear ownership
    - I don’t know

16. Who receives the results of AI red teaming exercises? *(Select all that apply)*
    - CIO
    - CISO
    - Risk Management/Internal Audit
    - CEO or other business executives
    - Board of Directors
    - Other internal stakeholders
    - Not reported
    - I don’t know

Footnotes and References

[1] Natasha Bajema, IEEE Spectrum, 'Why Are Large AI Models Being Red Teamed?', 15.05.2024. https://spectrum.ieee.org/red-team-ai-llms.

[2] Blake Bullwinkel, Amanda Minnich and others. Cornell University, Computer Science: Artificial Intelligence, eprint arXiv: 2501.07238, 'Lessons From Red teaming 100 Generative AI Products', 13 January 2025. https://arxiv.org/abs/2501.07238.

[3] Alice Qian Zhang, Ashlee Milton and others. CSCW Companion '24, arXiv:2407.07786v2, 'The Human Factor in AI red teaming: Perspectives from Social and Collaborative Computing', 9-13 November 2024. San Jose, Costa Rica. https://arxiv.org/pdf/2407.07786.

[4] Sadie Creese and Akshay Joshi, World Economic Forum, 'Securing innovation: A leader's guide to managing cyber risks from AI adoption', 21 January 2025. https://www.weforum.org/stories/2025/01/a-leaders-guide-to-managing-cyber-risks-from-ai-adoption/.

[5] Ibid.

[6]Louise Axon, Joanna Bouckaert and others. World Economic Forum, 'Artificial Intelligence and Cybersecurity: Balancing Risks and Rewards', January 2025. https://reports.weforum.org/docs/WEF_Artificial_Intelligence_and_Cybersecurity_Balancing_Risks_and_Rewards_2025.pdf.

[7] Holden Karnofsky, Carnegie Endowment for International Peace, 'Developing AI Risk Management With the Same Ambition and Urgency as AI Products', 16 December 2024, Washington DC, US. https://carnegieendowment.org/research/2024/12/developing-ai-risk-management-with-the-same-ambition-and-urgency-as-ai-products?lang=en.

[8] National Institute of Standards and Technology, U.S. Department of Commerce, 'Outline: Proposed Zero Draft for a Standard on AI Testing. Evaluation, Verification, and Validation', March 2025. https://www.nist.gov/document/outline-proposed-zero-draft-standard-ai-testing-evaluation-verification-and-validation.

[9] MITRE Corporation, 'The MITRE AI Maturity Model and Organizational Assessment Tool Guide', 3 November 2023. https://www.mitre.org/news-insights/publication/mitre-ai-maturity-model-and-organizational-assessment-tool-guide.

[10] MITRE Corporation, 'AI Red Teaming: Advancing Safe and Secure AI Systems', 31 July 2024. https://www.mitre.org/news-insights/publication/ai-red-teaming-advancing-safe-and-secure-ai-systems.

[11] MITRE Corporation, MITRE Atlas: ATLAS Matrix for AI Systems, https://atlas.mitre.org/.

[12] Tarleton Gillespie, Ryland Shaw and others. Cornell University, Computer Science: Computers and Society, eprint arXiv: 2412.09751, 'AI red-teaming is a sociotechnical problem: on values, labor, and harms', 7 January 2026. https://arxiv.org/abs/2412.09751.

[13] Alice Qian Zhang, Ashlee Milton and others. CSCW Companion '24, arXiv:2407.07786v2, 'The Human Factor in AI red teaming: Perspectives from Social and Collaborative Computing', 9-13 November 2024. San Jose, Costa Rica. https://arxiv.org/pdf/2407.07786.

[14] Blake Bullwinkel, Amanda Minnich and others. Cornell University, Computer Science: Artificial Intelligence, eprint arXiv: 2501.07238, 'Lessons From Red teaming 100 Generative AI Products', 13 January 2025. https://arxiv.org/abs/2501.07238.

[15] Subhabrata Majumdar, Brian Pendleton and Abhishek Gupta, Proceedings on Machine Learning Research 299: 1-20, 2025 Conference on Applied Machine Learning for Information Security, 'Red teaming AI Red teaming'. https://arxiv.org/pdf/2507.05538.

[16] Alberto Purpura, Sahil Wadhwa and others. Cornell University, Computer Science: Computation and Language, eprint arXiv: 2503.01742, 'Building Safe GenAI Applications: An End-to-End Overview of Red teaming for Large Language Models', 5 March 2025. https://arxiv.org/abs/2503.01742.

[17] Subhabrata Majumdar, Brian Pendleton and Abhishek Gupta, Proceedings on Machine Learning Research 299: 1-20, 2025 Conference on Applied Machine Learning for Information Security, 'Red teaming AI Red teaming'. https://arxiv.org/pdf/2507.05538.

[18] Natasha Bajema, IEEE Spectrum, 'Why Are Large AI Models Being Red Teamed?', 15.05.2024. https://spectrum.ieee.org/red-team-ai-llms.

[19] Alice Qian Zhang, Ashlee Milton and others. CSCW Companion '24, arXiv:2407.07786v2, 'The Human Factor in AI red teaming: Perspectives from Social and Collaborative Computing', 9-13 November 2024. San Jose, Costa Rica. https://arxiv.org/pdf/2407.07786.

[20] Alberto Purpura, Sahil Wadhwa and others. Cornell University, Computer Science: Computation and Language, eprint arXiv: 2503.01742, 'Building Safe GenAI Applications: An End-to-End Overview of Red teaming for Large Language Models', 5 March 2025. https://arxiv.org/abs/2503.01742.

[21] Natasha Bajema, IEEE Spectrum, 'Why Are Large AI Models Being Red Teamed?', 15.05.2024. https://spectrum.ieee.org/red-team-ai-llms.

[22] Tarleton Gillespie, Ryland Shaw and others. Cornell University, Computer Science: Computers and Society, eprint arXiv: 2412.09751, 'AI red-teaming is a sociotechnical problem: on values, labor, and harms', 7 January 2026. https://arxiv.org/abs/2412.09751.

[23] Shruti Srivastava, Kiranmazee and others. Cornell University, Computer Science: Cryptography and Security, eprint arXiv: 2602.21267, 'A Systematic Review of Algorithmic Red teaming Methodologies for Assurance and Security of AI Applications', 24 February 2026. https://arxiv.org/abs/2602.21267.

[24] Google, 'Responsible AI Progress Report', February 2026. https://ai.google/static/documents/ai-responsibility-update-2026.pdf

[25] Ibid.

[26] Subhabrata Majumdar, Brian Pendleton and Abhishek Gupta, Proceedings on Machine Learning Research 299: 1-20, 2025 Conference on Applied Machine Learning for Information Security, 'Red teaming AI Red teaming'. https://arxiv.org/pdf/2507.05538.

[27] Alice Qian Zhang, Jina Suh and others. Interactions: Forums, 'Effective Automation to Support the Human Infrastructure in AI Red teaming' 4 July – August 2025. https://interactions.acm.org/archive/view/july-august-2025/effective-automation-to-support-the-human-infrastructure-in-ai-red-teaming.

[28] Subhabrata Majumdar, Brian Pendleton and Abhishek Gupta, Proceedings on Machine Learning Research 299: 1-20, 2025 Conference on Applied Machine Learning for Information Security, 'Red teaming AI Red teaming'. https://arxiv.org/pdf/2507.05538.

[29] Ibid.

[30] OWASP, GenAI Security Project, 'GenAI Red Teaming Guide', January 22, 2025, https://genai.owasp.org/resource/genai-red-teaming-guide/.

[31] Subhabrata Majumdar, Brian Pendleton and Abhishek Gupta, Proceedings on Machine Learning Research 299: 1-20, 2025 Conference on Applied Machine Learning for Information Security, 'Red teaming AI Red teaming'. https://arxiv.org/pdf/2507.05538.

[32] Cloud Security Alliance, 'Agentic AI Red teaming Guide', 28 May 2025, https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide.