hackerone

SOLUTION BRIEF

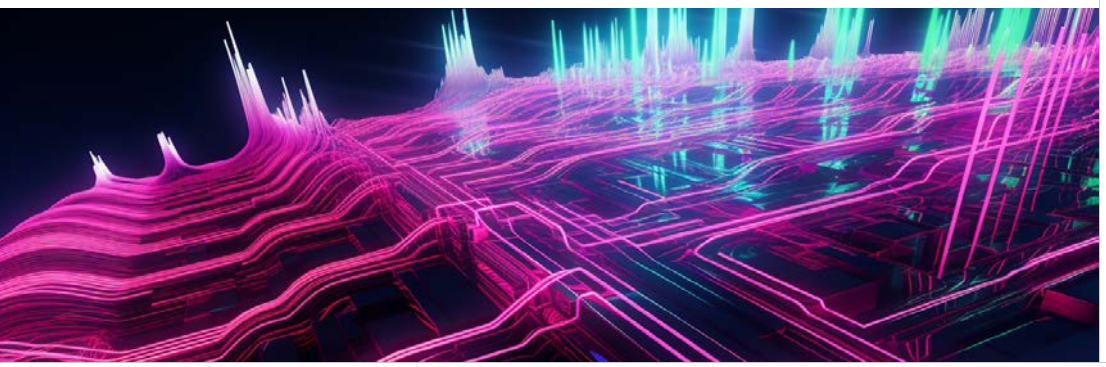
HackerOne AI Red Teaming

For Organizations Leading the Way in Safe and Secure AI Innovation

Targeted In-Depth Human Testing For AI Models

The swift adoption of AI introduces inherent safety and security risks that require preemptive strategies to prevent potential misuse. Establishing robust guiding principles for responsible AI implementation, alongside thorough and proactive security testing, is vital for mitigating technical vulnerabilities. These measures promote transparency and ensure adherence to ethical standards in AI deployment.

HackerOne's AI red teaming addresses the novel challenges of AI safety and security for businesses launching new AI deployments. This approach involves targeted offensive testing, harnessing the collective skills of ethical hackers proficient in AI and prompt hacking. It incentives the identification and remediation of critical vulnerabilities in your AI assets and strengthens your systems against potential risks, biases, and malicious exploits, assuring resilient and secure AI applications.



Key Outcomes

AI System Compliance and Trust

HackerOne ensures adherence to legal and ethical standards for AI systems, essential for preserving public trust and avoiding legal ramifications.

AI Vulnerability Detection and Response

Our human-led offensive testing rapidly uncovers and resolves AI model and technology vulnerabilities.

Financial Risk Mitigation

HackerOne's services help avoid costly legal penalties and reputational damage by minimizing risks associated with AI-related liabilities.

Testing for AI Safety and Security

AI Safety focuses on preventing AI systems from generating harmful content, such as bomb-making instructions or offensive language, upholding responsible AI use and ethical standards.

AI Security, on the other hand, concentrates on safeguarding AI systems from security vulnerabilities. This includes protecting user information and preserving the integrity and confidentiality of the AI infrastructure.

AI Security Testing Use Cases

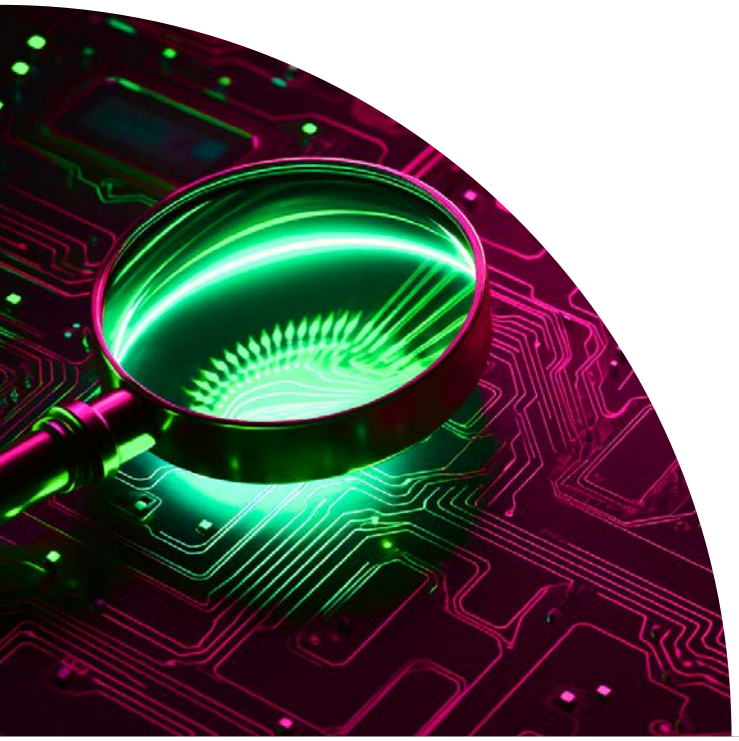
- **AI Implementation:** Our red teaming efforts are focused on testing and hardening AI implementation, such as AI chatbots communicating with APIs. When incorporating AI into applications, whether through custom models or off-the-shelf solutions, the process often presents complex challenges in areas like authorization and user input handling. These complexities are significant due to the natural language interfaces standard in AI systems. Identifying and effectively mitigating these risks is essential to ensure the security and functionality of the AI implementation.

AI Safety Testing Use Cases

- **Unsafe AI:** Unsafe AI can lead to harmful content generation by chatbots. It's crucial to guarantee the ethical deployment of AI systems and implement strict measures to avert threats to safety, loss of user trust, and brand reputation damage.
- **Malicious Use:** The malicious use of AI, especially in creating deceptive tools such as deepfakes and automated CAPTCHA solvers, represents a serious concern for users. Vigilant defenses are crucial to combat these worst-case scenarios, ensuring an AI deployment doesn't result in the enablement of misinformation, privacy infringements, and the abuse of advanced AI technologies.

HackerOne Red Teaming Capabilities

- **Strategic Flexibility:** Targeted vulnerability identification tailored to immediate security needs, enabling a custom engagement suited for your unique threat model or criteria. Get a strategic resource and skill allocation without long-term commitments.
- **Rapid Deployment:** The quick initiation of security testing programs to address urgent concerns, drawing on the community's expertise for swift, impactful assessment of crucial security areas.
- **Hybrid Talent Strategy:** A combination of AI/ML expertise with the unique perspectives of our diverse hacker community to uncover and resolve sophisticated vulnerabilities.
- **Intelligent Co-Pilot:** The use of Hai, our proprietary AI chatbot currently in development, to enrich vulnerability report analysis and improve dialogue with HackerOne's security researchers.



"As we see the technology mature and grow in complexity, there will be more ways to break it. We're already seeing vulnerabilities specific to AI systems, such as prompt injection or getting the AI model to recall training data or poison the data. We need AI and human intelligence to overcome these security challenges."

Katie Paxton-Fear, aka @InsiderPhD
Hacker specializing in AI

Securing AI With the World's Most Diverse Ethical Hacker Community

In HackerOne's skilled, global hacking community, over 750 active hackers specialize in prompt hacking and other AI testing methodologies. This robust and constantly growing community supports AI red teaming activities through various security testing engagement types offered by HackerOne, each designed to cater to specific aspects of AI safety and security:

Conduct continuous offensive testing through a **Bug Bounty Program**

Perform targeted hacker-led testing with a time-bound **Challenge**

Assess an entire application with a **Pentest** or **Code Security Audit**

Conduct **Spot Checks** for smaller, bite-sized AI features like chatbots, or predictive text inputs

---

"There are now suddenly a whole host of attack vectors for AI-powered applications that weren't possible before. Tricking the AI into doing or revealing something that it shouldn't. One of the reasons for AI security issues is that these applications are new and developing really fast; they don't necessarily take security very seriously."

Joseph Thacker, aka @rezO
Hacker specializing in AI

hackerone

Contact HackerOne to learn more about HackerOne's AI Red Teaming.

hackerone.com