bugcrowd

DATA SHEET

AI Safety & Security Solutions

Adopt AI with confidence after specialized AI Red Teaming from the crowd

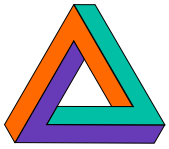
CHALLENGES

The adoption of LLM applications and other AI systems promises revolutionary competitive advantages, just as technologies like mobile apps, cloud computing, and IoT did in the past. However, as with any new technology wave, AI adds significant new vulnerabilities to the attack surface (some security-related and some safety-related) that few organizations fully recognize, with the risk often amplified by deep integration with other systems.

SOLUTION

The Bugcrowd Platform's AI Safety & Security portfolio meets the challenges of the moment. Our platform engages trusted hackers to use the same tools and processes threat actors do to probe your systems and find security or safety vulnerabilities on your behalf. You can then beat threat actors to the punch by patching up these vulnerabilities before they lead to damage. That helps meet regulatory requirements for AI red teaming and other security and safety standards described in Executive Order 14110, OMB Memorandum M-24-10, and the EU AI Act.

- ✓ **Uncover data bias and other hidden risks**
  Incentivized crowdsourcing engagements like AI Bias Assessments and bug bounties can uncover data bias and other vulnerabilities that traditional testing will miss.
- ✓ **Run targeted penetration tests on AI systems**
  AI Penetration Testing can deliver private, targeted, time-bound offensive testing to uncover hidden vulnerabilities in LLM and other AI applications.
- ✓ **Create a "neighborhood watch" for AI risk**
  Standing up a vulnerability disclosure program (VDP) gives the hacker/researcher community a formal way to altruistically report flaws in LLMs applications and other AI systems, before threat actors can find them.



As AI models and the security industry evolve together, AI will come to play three significant roles in the industry: tool, target, and threat.

AI as a Tool
Both sides of the security battlefield will use AI systems to scale up their attacks/defenses.

AI as a Target
Threat actors will exploit vulnerabilities in companies' AI systems.

AI as a Threat
AI models have been known to cause harm, including perpetuating biases or promoting hate speech.

For AI Security: **AI Penetration Testing**

Bugcrowd AI Pen Tests are designed to uncover common AI security flaws using a testing methodology based on our open-source Vulnerability Rating Taxonomy—which draws from the OWASP LLM Top 10 while adding other flaws reported by hackers on our platform. Issues found can include prompt injection, output handling, and data poisoning vulnerabilities.

- Trusted, vetted pentesters with specialized skills and experience in AI testing
- 24/7 visibility into timelines, findings, and pentester progress
- Equally effective on apps based on third-party or private models
- Detailed final report with remediation advice, as well as retesting to validate fixes

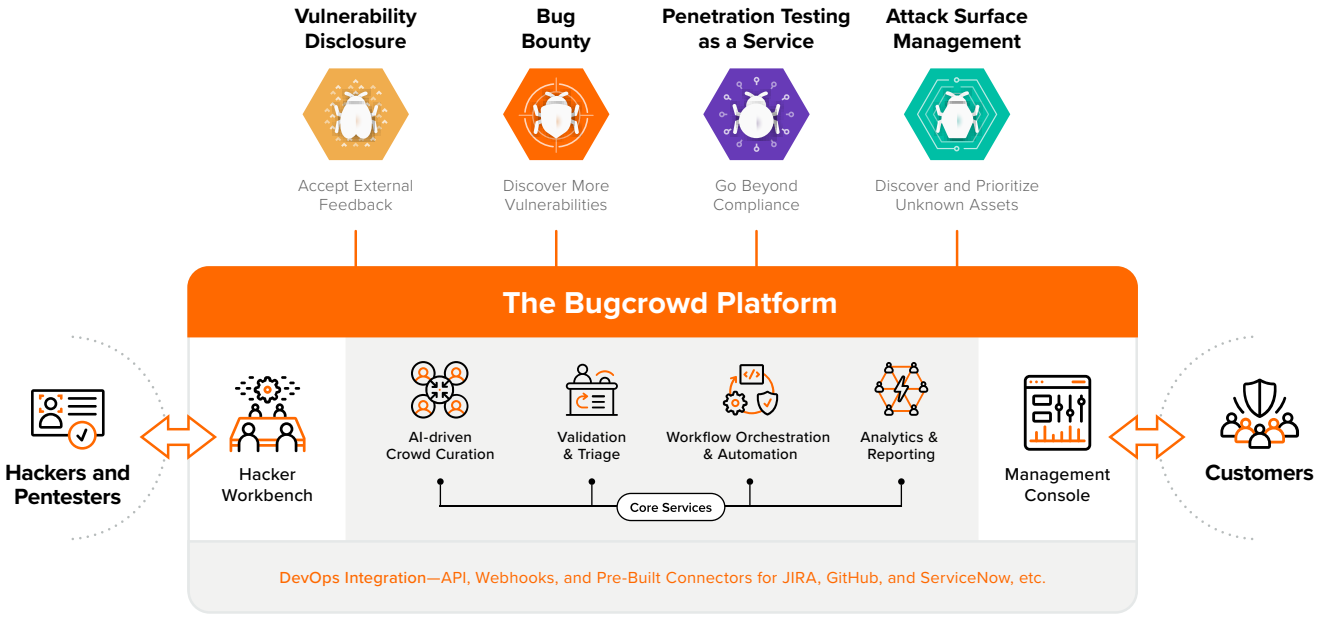
For AI Security: **AI Bias Assessments**

Bugcrowd AI Bias Assessments are safety-focused private engagements that activate trusted, 3rd-party security researchers to identify and prioritize data bias flaws in LLM applications. Participants are rewarded based on successful demonstration of impact, with more impactful findings earning higher payments.

- Detects numerous types of data bias in LLM apps that traditional testing will miss
- Streamlines and accelerates LLM adoption
- Contributes to compliance with AI Safety regulations for red teaming and bias (detection (eg EO 14410
- Builds relationships with AI specialists in the hacker/security researcher community

Why Bugcrowd

The Bugcrowd Platform helps customers defend themselves against cybersecurity attacks byconnecting with trusted, skilled hackers to take back control of the attack surface. Our AI-powered platform for crowdsourced security is built on the industry's richest repository of data about vulnerabilities and hacker skill sets, activating the ideal hacker talent needed on demand, and bringing scalability and adaptability to address current and emerging threats.



BEST SECURITY ROI FROM THE CROWD

We match you with trusted security researchers who are perfect for your needs and environment across hundreds of dimensions using machine learning.

INSTANT FOCUS ON CRITICAL ISSUES

Working as an extension of the platform, our global security engineering team rapidly validates and triages submissions, with P1s often handled within hours.

CONTEXTUAL INTELLIGENCE FOR BEST RESULTS

We apply over a decade of knowledge accumulated from experience devising thousands of customer solutions to achieve your goals for better outcomes.

CONTINUOUS, RESILIENT SECURITY FOR DEVOPS

The platform integrates workflows with your existing tools and processes to ensure that apps and APIs are continuously tested before they ship.

TESLA indeed credit karma Pinterest mastercard ATLASSIAN