hackerone

February 27, 2026

VIA EMAIL: secure.ai.infrastructure@dsit.gov.uk

Department for Science, Innovation and Technology (DSIT)
AI Security Institute (AISI)
National Cyber Security Centre (NCSC)

Re: Secure AI Infrastructure - Call for Information

Dear Sir or Madam,

HackerOne Inc. (HackerOne) submits the following comments in response to the UK government's call for information on secure AI infrastructure. We appreciate the opportunity to provide input and we commend the Department for Science, Innovation and Technology (DSIT), the AI Security Institute (AISI), and the National Cyber Security Centre (NCSC) for their joint leadership in advancing the development, assurance, and protection of next-generation AI systems.

By way of background, HackerOne is a global leader in offensive security solutions. Our HackerOne Platform combines AI with the ingenuity of the largest community of security researchers to find and fix security, privacy, and AI vulnerabilities across the software development lifecycle. The platform offers bug bounty, vulnerability disclosure, pentesting, AI red teaming, and code security. We are trusted by industry leaders like Amazon, Anthropic, Crypto.com, General Motors, GitHub, Goldman Sachs, Uber, and the U.S. Department of Defense.

Our comments focus on four practical and immediately impactful measures to strengthen secure AI infrastructure: structured AI red teaming, vulnerability disclosure programs (VDPs), bug bounty programs (BBPs), and strong protections for good-faith researchers.

1. **Encourage and Prioritize AI Red Teaming for Secure Systems**

AI systems are rapidly becoming embedded in critical infrastructure and high-impact services. While they offer significant economic and societal benefits, vulnerabilities in these systems can have systemic consequences. Security testing of AI is, therefore, both a business and national security imperative.

Traditional cybersecurity tools are not designed to identify the full range of AI-specific risks. AI systems introduce distinct vulnerability categories – including prompt injection, jailbreak techniques, model extraction, data poisoning, inference-based data leakage, and safety guardrail

bypass – that are not reliably detectable through conventional automated scanning. In addition, as AI systems scale, are fine-tuned, and are integrated with external tools and agentic components, the attack surface evolves dynamically and point-in-time assessments cannot keep pace with this change.

Structured AI red teaming addresses these gaps by engaging independent, diversely-skilled security researchers to adversarially test AI-enabled systems. Effective red teaming evaluates not only models in isolation, but also their integration into products, infrastructure, and external systems. When conducted prior to deployment and on an ongoing basis, red teaming strengthens security, safety, and resilience by identifying misuse pathways, integration flaws, and harmful behaviors that internal testing may miss.[1]

2. **Institutionalize Vulnerability Disclosure Programs (VDPs)**

Even comprehensive pre-deployment testing cannot identify every vulnerability. AI developers and deployers should implement clear, accessible VDPs that enable external researchers to responsibly report security flaws that surface post-deployment before they are exploited.

We strongly encourage AI developers and deployers to work with good-faith security and AI researchers to establish VDPs for receiving, triaging, and remediating disclosures related to AI security flaws prior to malicious exploitation. The UK can further support this ecosystem by promoting safe harbor protections that provide legal clarity for good-faith AI security research conducted within established program parameters. Transparent disclosure processes, standardized reporting expectations, and clear engagement rules strengthen accountability while protecting both operators and researchers.

HackerOne has developed a Good Faith AI Research Safe Harbor framework, an industry initiative designed to establish clear authorization and legal protections for researchers testing AI systems responsibly, and one that governments may look to as a model.[2] As AI capabilities scale rapidly across critical products and services, legal ambiguity surrounding testing can slow responsible research and inadvertently increase risk. A safe harbor framework reduces that friction by providing organizations with a shared standard that enables vulnerabilities to be identified and addressed more quickly and effectively.

3. **Scaling Bug Bounty Programs for AI Infrastructure**

Bug bounty programs build on VDPs by creating structured, incentive-based models for continuous security testing. This approach can and should be extended to AI-specific risks.

---

[1] HackerOne, AI Red Teaming Explained by AI Red Teamers, https://www.hackerone.com/blog/ai-red-teaming-explained-by-red-teamers.
[2] Hackerone AI Research Safe Harbor Statement, https://docs.hackerone.com/en/articles/13376522-ai-research-safe-harbor-statement.

Beyond technical security vulnerabilities, AI systems can introduce non-security harms, including problematic behavior, discrimination, safety failures, and other unintended outputs. For this reason, AI operators should consider adopting incentivized continuous testing programs. AI bug bounties apply the same incentive-based model used in security bug bounties, but include a focus on identifying and remediating AI-specific unintended outcomes. By incentivizing independent researchers to examine model outputs and system behavior, organizations can uncover algorithmic flaws that otherwise would go unnoticed.

This approach has demonstrated real-world value. HackerOne facilitated a public algorithm review in which the AI research community uncovered a number of validated unintended outcomes from social media algorithms.[3] The results of the engagement confirmed that the collaborative efforts of the broader research community can improve AI systems and reduce unintended outcomes.[4]

* * *

As AI systems scale across critical infrastructure, structured AI red teaming, formal vulnerability disclosure mechanisms, incentive-based testing models, and robust protections for good-faith researchers are essential to ensuring security, resilience, and trustworthiness.

HackerOne welcomes continued collaboration with DSIT, AISI, and NCSC and stands ready to support the UK's efforts to operationalize secure AI infrastructure and reinforce its position as a trusted leader in responsible AI development. If we can be of further assistance, please contact the HackerOne Policy Team at policy-team@hackerone.com.

Respectfully,

Ilona Cohen
Chief Legal and Policy Officer
HackerOne

[3] Twitter, Twitter's First Algorithmic Bias Bounty Challenge
https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge
[4] Twitter, Learnings from Twitter's Algorithmic Bias Bounty
https://blog.twitter.com/engineering/en_us/topics/insights/2021/learnings-from-the-first-algorithmic-bias-bo unty-challenge