hackerone

August 9, 2024

VIA ELECTRONIC SUBMISSION

AI Cyber Security Call for Views
Secure Code and Standards
Cyber Security & Digital Identity Directorate
Department for Science, Innovation & Technology
Level 4
100 Parliament Street
Westminster, London
SW1A 2BQ

Re: Call for Views on the cyber security of AI

Dear Viscount Camrose,

HackerOne Inc. (HackerOne) submits the following comments in response to the Department for Science, Innovation & Technology's (DSIT) Call for Views on the Cyber security of AI.[1] HackerOne appreciates the opportunity to provide input, and we commend DSIT for its openness in working with industry stakeholders on this important issue.

HackerOne is the global leader in human-powered security. We leverage human ingenuity to pinpoint the most critical security flaws across your attack surface to outmatch cybercriminals. HackerOne's Attack Resistance Platform combines the most creative human intelligence with the latest artificial intelligence to reduce threat exposure at all stages of the software development lifecycle. From meeting compliance requirements with pentesting to finding novel and elusive vulnerabilities through bug bounty, HackerOne's elite community of ethical hackers helps organisations transform their businesses with confidence. HackerOne has helped find and fix vulnerabilities for sector leaders including Coinbase, General Motors, GitHub, Goldman Sachs, the Financial Times, Starling Bank, the U.S Department of Defense, and the UK Ministry of Defence.

HackerOne's comments focus on the following principles:

- Principle 1: Raise staff awareness of threats and risks
- Principle 6: Secure your infrastructure
- Principle 9: Conduct appropriate testing and evaluation

[1] Department for Science, Innovation & Technology (DSIT), *A call for views on the cyber security of AI*, 18 July 2024, https://www.gov.uk/government/calls-for-evidence/cyber-security-of-ai-a-call-for-views/a-call-for-views-on-the-cyber-security-of-ai.

- Principle 11: Maintain regular security updates for AI model and systems

Principle 1: Raise Staff Awareness of Threats and Risks

HackerOne agrees that organisations should establish security-awareness training to educate their staff on the evolving threat landscape specific to AI systems. Employees at all levels should be aware of the latest vulnerabilities that could impact their systems and AI security best practices. Such training should include information about both security and non-security vulnerabilities. Specifically, we recommend that DSIT amends provision 1.3 to include training on potential AI harms, including discrimination, bias, and security risks. A comprehensive approach to training will empower employees to contribute effectively to the organisation's security efforts and enhance the overall resilience and ethical deployment of AI systems.

Principle 6: Secure your infrastructure

All organisations would benefit from the implementation of Vulnerability Disclosure Programs (VDPs). These programs are increasingly recognised in international standards and regulations, such as the Federal Cybersecurity Vulnerability Reduction Act introduced last year in the U.S. Congress, which would mandate all U.S. government contractors to implement a VDP.[2] HackerOne commends the inclusion in the Code of Practice of a requirement for developers and system operators to “implement and publish an effective vulnerability disclosure process to support a transparent and open culture.” We further encourage DSIT to extend that provision 6.3 requirement to ensure that the process can address both security vulnerabilities and non-security AI flaws. By addressing both security vulnerabilities and non-security flaws, organisations can ensure a more comprehensive approach to risk management and resilience of their AI systems.

In addition, developers and system operations should consider additional vulnerability management practices beyond VDPs. For example, we recommend the inclusion of references to bug and bias bounty and penetration testing programs. Bug bounties provide an additional layer of security by incentivizing external researchers to identify and mitigate security vulnerabilities, while bias bounties offer similar benefits in regards to AI trustworthiness flaws like bias. Penetration testing simulates real-world attacks to uncover vulnerabilities before adversaries can exploit them. These proactive approaches enhance the effectiveness of vulnerability management and foster a collaborative environment for improving AI system security and trustworthiness.

Principle 9: Conduct appropriate testing and evaluation

HackerOne strongly supports the inclusion of the requirement for validation and assessment of AI models. While it has become common to generally describe risk management

[2] H.R.5225, Federal Cybersecurity Vulnerability Reduction Act, https://www.congress.gov/118/bills/hr5255/BILLS-118hr5255ih.pdf

practices for AI models, this is often presented without external validation. HackerOne believes that independent external evaluation is essential for ensuring that AI models are thoroughly tested and validated by unbiased parties, leading to more robust, secure, and trustworthy AI systems.

We also support provision 9.2.2 which states that evaluations should include "red-teaming or other adversarial testing." HackerOne encourages further defining red-teaming to include both adversarial and non-adversarial testing, addressing both security and non-security risks. This aligns with emerging consensus, as represented by the U.S. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, which describes red-teaming as a structured testing effort to identify both security vulnerabilities and non-security flaws, including bias, discrimination and other harmful outputs.[3] To promote alignment with emerging best practice, we urge the UK to consider this change.

Principle 11: Maintain regular security updates for AI model and systems

Finally, while we support maintaining regular security updates and patches in AI systems, we have concerns about provision 11.2.1, which requires developers to "publish bulletins in response to vulnerability disclosures, including detailed and complete Common Vulnerability Enumeration (CVE) information in cases where updates can't be provided." We encourage DSIT to clarify that this provision does not mandate the release of detailed CVE information even when not yet mitigated. Without the clarification, such an interpretation might inadvertently lead to the disclosure of unmitigated vulnerabilities or flaws to third parties or customers.

Conclusion

HackerOne appreciates the opportunity to provide a response to this call for views. As the conversation around this topic continues to evolve, we would welcome the opportunity to further serve as a resource and provide insights on how to raise the standard for cyber security in AI.

* * *

Respectfully Submitted,

Ilona Cohen
Chief Legal and Policy Officer
HackerOne

[3] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 3(d), Oct. 30, 2023, www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trust worthydevelopment-and-use-of-artificial-intelligence.