hackerone

March 14, 2025

VIA ELECTRONIC SUBMISSION
Regulations.gov

Re: Comments on the Development of an Artificial Intelligence (AI) Action Plan

Dear Sir or Madam,

HackerOne Inc. ("HackerOne") submits the following comments in response to the Office of Science and Technology Policy (OSTP) and Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO)'s Request for Information ("RFI") on the Development of an Artificial Intelligence (AI) Action Plan.[1] HackerOne appreciates the opportunity to provide input on this important issue.

By way of background, HackerOne is a global leader in finding and fixing critical vulnerabilities and AI security issues. Our industry-leading HackerOne Platform combines AI with the expertise of the world's largest community of security researchers to uncover and remediate vulnerabilities and AI security issues across the software development lifecycle. The platform offers bug bounty, vulnerability disclosure, pentesting, code review, and AI red teaming.

HackerOne consistently advocates for widespread adoption of cybersecurity measures that have proven effective at addressing unmitigated vulnerabilities in both commercial and government contexts. This advocacy extends to the realm of AI, where we set up bug bounties for AI security testing, and we also conduct algorithmic reviews to help reduce unintended outcomes. As the White House works to define its priorities around AI, our recommendations focus on the importance of red teaming and leveraging the hacker community to help secure and test AI systems.

Encourage and prioritize AI Red Teaming for secure systems

AI is poised to usher in transformative changes that can supercharge the U.S. economic productivity and improve daily life of all Americans. However, this immense potential comes with responsibilities to ensure that AI systems are secure. Testing AI systems for security is more than just a best practice - it is a business and national security imperative. Red teaming and other structured tests can proactively identify weaknesses that automated or internal scanning may miss, reducing the risk that threat actors will circumvent system safeguards.

To that end, we strongly encourage the government to:

---

[1] Office of Science and Technology Policy (OSTP) and Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), Request for Information, Development on an Artificial Intelligence (AI) Action Plan, Feb. 6, 2025, https://www.federalregister.gov/documents/2025/02/06/2025-02305/request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan

1. **Incentivize red teaming for higher risk AI systems.** The government should incentivize red teaming for AI systems that perform key functions for critical infrastructure, or that could cause harm to individuals, public health or safety, or national economic or security interests. This red teaming should occur prior to deployment of the AI system, as well as on a periodic basis to help ensure system security is maintained.

2. **Update federal safe harbors for sharing AI red teaming results.** The government should ensure legal frameworks for security information sharing are adapted to encourage and protect information sharing for unintended outputs in AI systems. This includes the Cybersecurity Information Sharing Act of 2015, Section 1201 of the Digital Millennium Copyright Act, and ensuring AI model deployers have methods for receiving disclosures about model flaws from independent sources and the general public. The government should further ensure that independent good faith AI researchers are protected from legal risk for finding and responsibly disclosing AI vulnerabilities and flaws by, for example, exempting good faith AI research from copyright law restrictions[2] and extending U.S. Department of Justice guidance that protects good faith security research from prosecution to also protect good faith AI research.[3]

3. **Support research and development in AI red teaming.** The government should invest in R&D to advance AI red teaming methodologies. This includes promoting AI red teaming test beds, developing more sophisticated tools for testing AI models and validating results, and expanding the availability of skilled professionals who can conduct AI red teaming. Supporting R&D in AI red teaming can help the U.S. maintain a competitive edge and a high degree of resilience in AI.

Strengthening AI security and reducing unintended outputs

American leadership in AI development is bolstered through demonstrating that models protect consumers' privacy and reduce the risk of unintended consequences. It is critical for organizations to establish processes to receive and respond to disclosures of information about AI security and flaws from external sources. HackerOne encourages the government to work with good faith security and AI researchers to establish robust processes for receiving and responding to disclosures about AI security and flaws. NIST should enable coordinated flaw disclosures with a safe harbor, since the interconnected nature of AI systems means flaws can transfer across platforms. A unified disclosure process is essential, with standardized AI flaw reports and clear engagement rules to ensure a timely, transparent resolution that protects AI researchers.

---

[2] HackerOne, *Short Reply Comment to US Copyright Office, Ninth Triennial Section 1201 Proceeding (2024), Class 4: Computer Programs–Generative AI Research*, Mar. 18, 2024, https://www.copyright.gov/1201/2024/comments/reply/Class%204%20-%20Reply%20-%20HackerOne%20Inc..pdf.

[3] HackerOne, *Request to Develop a Charging Policy under the CFAA to Protect Independent AI Trustworthiness Research*, Apr. 16, 2024, https://www.hackerone.com/sites/default/files/2025-02/HackerOne-Letter-to-DOJ-re-AI-Testing.pdf.

hackerone

Providing incentives for good faith researchers to identify system weaknesses will better empower organizations to leverage the expertise of the security and AI communities to secure models against unintended outputs. These external experts bring varied backgrounds and knowledge that enable them to detect weaknesses and flaws that may go unnoticed by a homogenous internal team, providing a more comprehensive analysis than internal testing alone. Independent third-party evaluations provide broader and more continuous scrutiny, helping identify risks that may otherwise go unnoticed. Drawing from the success of vulnerability disclosure programs (VDPs) in cybersecurity, adopting a similar “default to disclosure” approach for flaws identified by these independent evaluators in AI systems can improve security and risk management. VDPs have proven effective in identifying and addressing vulnerabilities in software, and a similar mindset for AI models can help ensure that flaws are quickly recognized and addressed. Encouraging transparency around AI models and flaws, along with supporting rigorous third-party evaluations, enables stakeholders to proactively identify and address risks, benefiting the entire AI ecosystem. By incentivizing the discovery and disclosure of AI model flaws, the government and organizations can build public trust and foster ecosystem-wide improvements in AI development.

* * *

HackerOne appreciates the opportunity to provide comments to this request for information. As the conversation around this topic continues to evolve, we would welcome the opportunity to further serve as a resource and ensure the security of AI.

Respectfully submitted,

Ilona Cohen
Chief Legal and Policy Officer
HackerOne