hackerone

March 14, 2025

VIA ELECTRONIC SUBMISSION
CyberAIProfile@nist.gov

Re: NIST Cybersecurity and AI Workshop Concept Paper

Dear Sir or Madam,

HackerOne Inc. (HackerOne) submits the following comments in response to the National Institute of Standards and Technology's (NIST) Cybersecurity and AI Workshop Concept Paper.[1] We thank NIST for the opportunity to provide input on this important proposal.

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 help reduce undesirable outputs in AI. As the demand for secure AI grows, HackerOne is best positioned to assist enterprises in navigating the complexities of deploying AI models responsibly.

We believe that our expertise in ethical hacking and cybersecurity, coupled with our active engagement in AI security, positions HackerOne as a key contributor to the conversation on the intersection of cybersecurity and AI. We recognize the dual role that AI plays in both enhancing and complicating security measures, and we are eager to collaborate on developing standards that promote resilience in AI systems.

Scope and Focus

We support NIST's efforts to develop actionable guidance on managing cybersecurity and AI risks. The NIST Cybersecurity Framework (CSF) has been an invaluable tool for organizations in navigating cybersecurity challenges, aligning risk management strategies with business goals, and improving resilience. Given the growing convergence of AI and cybersecurity, we believe that the introduction of an AI profile within the CSF would be highly beneficial. Building upon the well-established NIST CSF would encourage widespread adoption of best practices for AI systems and help organizations better understand, assess, and manage both the risks and opportunities associated with AI.

[1] National Institute of Standards and Technology's (NIST) Cybersecurity and AI Workshop Concept Paper, Feb. 14, 2025, https://www.nccoe.nist.gov/sites/default/files/2025-02/cyber-ai-concept-paper.pdf.

While we support the three primary focus areas outlined in the Concept Paper – securing AI system components, thwarting AI-enabled attacks, and leveraging AI in cybersecurity approaches – we believe these areas could benefit from further elaboration and specific guidance. In particular, we recommend that the following considerations be incorporated to further strengthen these focus areas:

Strengthening AI Risk Management

To effectively prevent AI-related attacks and address the cybersecurity risks associated with AI, organizations must adopt targeted practices designed to identify and mitigate vulnerabilities both in the design and operational deployment of AI systems. We recommend that NIST encourage the adoption of various AI testing approaches and methods as core components of cybersecurity strategies and business risk management frameworks.

a) Incorporate AI Red Teaming into Practices

HackerOne encourages NIST to emphasize the integration of AI red teaming into an organization's cybersecurity practices. AI red teaming is essential for uncovering how adversaries might exploit vulnerabilities in AI systems, providing valuable insights that help organizations strengthen their defenses and improve overall resilience.

We recommend that NIST begin by clearly defining AI red teaming, outlining its key processes and methodologies. This definition should include the purpose of the testing, the general approach to conducting tests, the specific goals or outcomes sought, and the relevant metrics used to evaluate performance. A well-defined framework would allow organizations to understand how AI red teaming fits into their broader risk management strategies and help bring clarity to the various types of testing involved. While the primary focus for NIST's new guidance is on cybersecurity and AI, it is important to organizations that AI red teaming goes beyond traditional cybersecurity concerns. Their frameworks should also address broader security risks and unintended outputs, including potential chemical, biological, radiological, and nuclear (CBRN) threats. By adopting a more holistic approach that includes these wider security issues, organizations can ensure they are prepared for a wide range of potential risks that AI systems may introduce.

Additionally, such a framework should emphasize the importance of third-party evaluations as part of a holistic approach to AI red teaming. In-house evaluations, while common, can sometimes miss critical flaws, and AI evaluation infrastructure is still evolving. Independent third-party evaluations provide broader and more continuous scrutiny, helping identify risks that may otherwise go unnoticed.

As NIST looks to define AI red teaming, we encourage the inclusion of best practices drawn from real-world engagements. Through our collaborations with leading technology companies to evaluate AI deployments for unintended outcomes, HackerOne has developed a comprehensive

and evolving playbook for AI red teaming.[2] We recommend that NIST incorporate these principles into its guidance on AI red teaming and testing. Key considerations from HackerOne's playbook include:

- Team Composition: Diversity in background, experience, and skill sets is pivotal for ensuring secure AI.
- Collaboration and Size: Collaboration among AI red teaming members holds unparalleled significance, and a sufficient number of testers, based on the duration and scope of the engagement, should be engaged to enable the benefits of collaboration across diverse and global perspectives.
- Duration: Because AI technology is evolving so quickly, engagements between 15 and 60 days have worked best to assess specific aspects of AI, but a continuous engagement without a defined end date can be most effective when the systems in scope are rapidly changing.
- Context and Scope: Unlike much traditional security testing, AI red teamers cannot approach a model blindly. Providing both broad context and specific scope is crucial to determining the AI's purpose, deployment environment, existing safety features, and limitations. Contextual information should include diagrams of the AI deployment and the AI system's expected data access and actions.
- Clear objectives: Clear and precise objectives are needed for efficient AI testing and red teaming for unintended outcomes. For example, an objective like "Generate image of [harmful subject matter]" is too vague and may result in false positive reports that technically meet the letter of the objective but not the intention. A clearer objective would be "Generate image of [harmful subject matter] that includes [details of harmful subject matter]."
- Incentive Model: Tailoring the incentive model is a critical aspect of the AI testing playbook. A hybrid economic model that includes both fixed-fee participation rewards in conjunction with rewards for achieving specific outcomes (akin to bounties) has proven most effective.

b) Incorporate Vulnerability Disclosure Policies and Bug Bounties for AI

Additionally, we recommend that NIST encourage the integration of Vulnerability Disclosure Policies (VDPs) and Bug Bounty Programs (BBPs) in the context of AI systems. AI technologies, due to their inherent complexity and dynamic nature, often generate unpredictable or unintended outcomes that may not be immediately evident.

To address these potential risks in a timely and efficient manner, organizations must be equipped with structured processes for discovering, triaging, and mitigating AI-related vulnerabilities. The integration of VDPs and BBPs provides a robust mechanism for leveraging

[2] HackerOne, "An Emerging Playbook for AI Red Teaming With HackerOne," https://www.hackerone.com/thought-leadership/ai-safety-red-teaming.

the collective expertise of the security research community to uncover these vulnerabilities before they can be exploited or cause harm. NIST should issue guidance that promotes a unified disclosure process, with standardized AI flaw reports and template clear engagement rules to help ensure a timely, transparent resolution that protects independent AI researchers. A related effort could include the creation of a centralized disclosure hub to aggregate and analyze AI flaws, which would promote transparency, enable sharing across organizations, and enhance the overall security of AI models.[3]

* * *

Overall, we believe that NIST's guidance should note the value of both AI-driven testing tools and human testers, recognizing that human red teamers and security experts bring unique insights that complement the capabilities of automated systems. HackerOne appreciates the opportunity to contribute to this important initiative and looks forward to further engaging with NIST in developing robust, effective cybersecurity standards for AI. We would be glad to serve as a resource as NIST continues its work.

Respectfully Submitted,

Ilona Cohen
Chief Legal and Policy Officer
HackerOne

[3] *See, e.g.,* Longpre, et al., In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI (2025), https://drive.google.com/file/d/1nY22xxJVqi4_ZcyhBIvBxsmNBObAW08s/view.