hackerone

February 2, 2024

VIA ELECTRONIC SUBMISSION

National Institute of Standards and Technology
100 Bureau Drive, Stop 2000
Gaithersburg, MD 20899

Re: RFI Related to NIST's Assignments Under the Artificial Intelligence Executive Order

HackerOne submits the following comments in response to the Request for Information (RFI) related to National Institute of Standards and Technology (NIST)'s responsibilities under Sections 4.1, 4.5 and 11 of the recent Artificial Intelligence (AI) Executive Order (EO) 14110.[1] We thank NIST for the opportunity to provide input on this important proposal.

By way of background, HackerOne pinpoints the most critical security flaws across an organization's attack surface with continual adversarial testing to outmatch cybercriminals. HackerOne's Attack Resistance Platform blends the security expertise of ethical hackers with asset discovery, continuous assessment, and process enhancement to reduce threat exposure and empower organizations to transform their business with confidence.

HackerOne consistently advocates for widespread adoption of hacker-powered 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 and safety testing, and we also conduct algorithmic bias reviews to help reduce biases and undesirable outputs in AI. As the demand for secure and ethical AI grows, HackerOne is best positioned to assist enterprises in navigating the complexities of deploying AI models responsibly.

For the purposes of these comments, we focus on testing and "AI red teaming." In our comments, we highlight the importance of red teaming as an important security activity, but note that it's critical not to substitute it for other other safeguards, such as vulnerability management practices.

HackerOne recommends:

1. Develop a Technical Guide to AI Testing and Assessment.
2. Incorporate coordinated disclosure processes and bias bounties for harms unrelated to security and safety.

[1] 88 FR 88368, https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the.

Develop a technical guide to AI testing and assessment

EO 14110 requires NIST to develop guidelines for a variety of AI tests to enable deployment of safe, secure, and trustworthy systems. Though the EO and NIST's NPRM refer to these tests generically as "AI red teaming," there are several AI risk identification and assessment techniques that would benefit from NIST guidance. Through our partnership with leading technology firms to evaluate AI deployments for issues like safety, harmful output, and bias, HackerOne has developed an evolving playbook for a variety of AI red teaming engagements.[2] We suggest NIST incorporate these principles into its guidance for AI red teaming and testing. Some of the primary considerations in HackerOne's playbook include:

- Team Composition: Diversity in background, experience, and skill sets is pivotal for ensuring safe 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 safety, 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.
- Clear objectives: Clear and precise "degradation objectives" are needed for efficient AI testing and red teaming for bias and other non-security risks. 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]."
- Private vs. Public: While most AI red teaming engagements operate in private due to the sensitivity of safety issues, there are instances where public engagement - such as Twitter's algorithmic bias bounty challenge, facilitated by HackerOne - have yielded significant success. In general, the ecosystem benefits from transparency in security and safety and we encourage organizations to publish statistics regarding vulnerabilities and

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

algorithmic flaws that they find and fix, to the extent possible. That way there is greater visibility into the threat landscape and effectiveness of mitigation methods, and other organizations can use these learnings to become more resilient.

- 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 safety outcomes (akin to bounties) has proven most effective.

- Empathy and Consent: As surfacing AI-related issues may involve encountering harmful and offensive content, it is important to seek explicit participation consent from adults (18+ years of age), offer regular support for mental health, and encourage breaks between assessments.

We suggest NIST issue a high-level technical guide that describes major AI testing and examination processes and procedures. The guide should include the purpose of the testing, the general methodology for conducting the tests, the goals or outcomes sought by the tests, and relevant metrics for evaluating performance.

This format would help organizations determine where specific tests fit in broader organizational risk management efforts. It would also help bring consistency and clarity to terminology so that different tests are not conflated with one another. Lastly, it would help set broader baseline expectations for what a robust AI testing program should incorporate.

Incorporate coordinated disclosure processes and bias bounties for harms unrelated to security and safety

It is critical for organizations to establish processes to receive and respond to disclosures of information about AI security and flaws from external sources. Just as coordinated vulnerability disclosure and bug bounties are commonplace practices for security risks, coordinated disclosure and bias bounty processes are growing in adoption for non-security risks. These practices should be incorporated into AI red teaming guidance, technical guides to AI testing and assessment, and baseline AI risk management practices, and existing enabling frameworks such as liability protections should be extended to such activities.

Coordinated disclosure and bias bounties are processes that AI developers can utilize to proactively identify and address potential risks in AI systems. Coordinated vulnerability disclosure in security have helped companies become aware of vulnerabilities in their products and quickly remediate them prior to potential exploitation. Because AI systems must balance a wider variety of risks, AI system operators should incorporate disclosure and handling processes for non-security flaws. However, these processes may involve different personnel to evaluate and mitigate flaws rather than security teams because the skillset is distinct. Establishing a comprehensive disclosure and handling process separate from security vulnerability disclosure programs is helpful for effectively analyzing and responding to AI disclosures unrelated to security and safety.

AI operators can also adopt bias bounties, which are incentive programs that aim to discover and rectify biases in AI systems, focusing on uncovering and addressing instances of unfairness, discrimination, or unintended bias within algorithms. Although different from security bug bounties, which use the hacking community to identify and remediate vulnerabilities, bias bounty programs can rely on the hacking community to identify and collect algorithm flaws. This service is increasingly used by AI system operators to catch flaws that their internal evaluation regimes may have overlooked. For example, back in 2021, HackerOne facilitated a public bias review and the ethical hacker community played a crucial role in uncovering unexpected unfairness, discrimination, and unintended bias in Twitter algorithms.[3] The results of the engagement confirmed that our understanding of bias in AI can be improved when diverse perspectives and collaborative efforts are incorporated.[4]

We encourage NIST to acknowledge the variety of AI testing approaches and methods, and provide guidance that highlights the utility of different types of AI testing depending on the circumstances and particularities of AI systems, including in-house testing, penetration testing, and bug and bias bounties.

Conclusion

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 safety, security and reliability of AI.

*　　　*　　　*

Respectfully Submitted,

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-bounty-challenge