hackerone

September 9, 2024

VIA ELECTRONIC SUBMISSION

Re: NIST AI 800-1: Managing Misuse Risk for Dual-Use Foundation Models

Dear Sir or Madam,

HackerOne Inc. (HackerOne) submits the following comments in response to National Institute of Standards and Technology (NIST) AI Safety Institute's public guidance on Managing Misuse Risk for Dual-Use Foundation Models (NIST AI 800-1).[1] HackerOne appreciates the opportunity to provide input, and we commend NIST for its openness in working with industry stakeholders on this important issue.

HackerOne is the global leader in human-powered security, harnessing the creativity of the world's largest community of security researchers with cutting-edge AI to protect digital assets. The HackerOne Platform combines the expertise of our elite community and the most up-to-date vulnerability database to pinpoint critical security flaws across your attack surface. Our integrated solutions, including bug bounty, pentesting, code security audits, spot checks, and AI red teaming, ensure continuous vulnerability discovery and management throughout the software development lifecycle. HackerOne has helped find and fix vulnerabilities for leading companies in many industry sectors across the globe.

I. Scope

As adoption of AI continues to grow, we support NIST's efforts to build upon Executive Order 14410 and establish guidelines for holistic evaluation of AI models and managing misuse risks. However, we encourage NIST to clarify the scope of the draft guidelines. Currently, while NIST frames the guidelines as encompassing "safety, security, and trustworthiness," the scope is limited to "public safety," excluding key risks such as bias, discrimination, and reliability—essential aspects of "trustworthiness" as outlined in NIST's AI Risk Management Framework.[2]

[1] National Institute of Standards and Technology (NIST), NIST AI 800-1, Managing Misuse Risk for Dual-Use Foundation Models, Jul. 2024, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf.
[2] NIST, Artificial Intelligence Risk Management Framework, NIST AI 100-1, Jan. 2023, pg. 12, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

We recommend NIST clarify which specific trustworthiness characteristics and risks are included, and which are excluded, and revise the framing of the guidelines to avoid confusion of the scope. To the extent that other trustworthiness characteristics, such as fairness and reliability, are excluded, we suggest that NIST draft guidelines on managing misuse risk in relation to those characteristics as well, consistent with the direction of Executive Order 14410 to develop guidelines for trustworthiness (which is often included under the umbrella term "safety" by many in the industry).[3]

II. Practice 3.1: Assess the risk of model theft from relevant threat actors

HackerOne supports NIST's focus on assessing the risk of model theft. Protecting AI models from unauthorized access is critical to preventing threat actors from recreating and misusing the model. We agree with Practice 3.1's recommendation that developers should "consider using cybersecurity red teams and penetration testing to assess how difficult it would be for an actor to circumvent security measures."[4]

This recommendation aligns with industry best practices – simulating real-world attack scenarios is crucial for identifying potential weaknesses that could be exploited by adversaries. The fundamental testing methods recommended by NIST will help uncover vulnerabilities that static security measures may miss, providing critical insights for assessing the risk of model theft.

HackerOne urges NIST to make clear that red team and penetration testing should occur periodically, not only as a pre-deployment assessment. The effectiveness of safeguards can change over time in response to technological or product changes, including the development of new attack techniques, and continual assessment is needed to maintain a sufficient degree of protection against model theft. This aligns with NIST's recommendations under Practice 6.1, suggesting continual assessment of safeguards intended to detect misuse.

III. Practice 4.2: Use red teams to assess whether threat actors could bypass model and system safeguards and misuse any capabilities of concern.

HackerOne strongly agrees with NIST's recommendation to leverage red teams to assess the sufficiency of safeguards against misuse. 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.

---

[3] Executive Order on the Safe, Secure and Trustworthy Development of Artificial Intelligence, Section 4.1(a)(ii), Oct. 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[4] NIST AI 800-1, Managing Misuse Risk for Dual-Use Foundation Models, pg. 8.

HackerOne further agrees with the recommendation under Practice 4.2, which encourages developers to "consider providing red teams with available legal protections for their tasks, such as waiving terms of service and indemnifying them for legal liability for their interactions with the model." [5] These legal protections are essential for fostering an environment where red teams can operate without fear of legal repercussions, enabling them to thoroughly test systems for vulnerabilities. However, these protections should not be optional, and HackerOne urges NIST to remove the word "consider" to ensure red teams have the liability protections necessary for effective testing that leads to stronger AI model security.

IV. Practice 6.4: Provide safe harbors for third-party safety research

HackerOne strongly supports Practice 6.4's goal of providing safe harbors for independent good faith AI research. Providing legal protections for third-party researchers is essential to fostering an environment where security and non-security flaws can be identified and addressed proactively. However, below are a few recommendations to enhance this practice.

The first recommendation under Practice 6.4 suggests "publishing a clear vulnerability disclosure policy for model safety issues that outlines how such vulnerabilities should be shared with the developer and the public, and how the organization will respond to reported vulnerabilities."[6] HackerOne recommends against limiting the policy to"model safety issues." The policy should address not only model safety issues but also security, safety, and other trustworthiness concerns, which reflect the full scope of misuse risk and the likely variety of submissions from third party researchers.

Additionally, while Practice 6.4's first recommendation refers to sharing vulnerabilities with the public, we would caution organizations regarding the importance of confidentiality in this process. Whenever possible, vulnerabilities should be reported directly to the developer to allow for remediation before any public disclosure, and organizations should avoid publicly disclosing vulnerabilities until adequate remediation measures have been implemented or if there is a significant risk that warrants immediate public awareness. This reduces the risk that threat actors will successfully exploit the newly disclosed vulnerabilities, and maintaining confidentiality fosters trust between researchers and developers. This approach is consistent with best practices for security vulnerability disclosure and handling processes.

We agree with the second recommendation under practice 6.4 to "publish a safe harbor policy that commits to not pursuing legal action against or restricting the accounts of external safety researchers that act in good faith and comply with the vulnerability disclosure policy."[7]

[5] *Id.*, pg 18.
[6] *Id.*, pg 15.
[7] *Id.*

Without clear protections in place, good faith researchers are susceptible to legal threats, which can deter them from reporting vulnerabilities due to fear of potential legal repercussions. By establishing this safe harbor policy, organizations can encourage more researchers to disclose any critical system weaknesses to the developer in a timely fashion. Such a policy not only helps protect researchers but also strengthens the security of AI model developers and the overall digital ecosystem.

Finally, we appreciate the consideration of "providing support and accommodations for vetted external researchers, such as access to models with fewer safeguards for conducting post-deployment red-teaming exercises."[8] Providing such access helps external experts provide comprehensive assessments and ensure that potential models are thoroughly tested at crucial points in their development.

V. Practice 6.5: Create bounties for issues related to the misuse risk

HackerOne agrees on the importance of establishing robust processes for receiving and responding to disclosures about AI security and flaws from external sources. 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 misuse risk.These external experts bring varied background and knowledge that enable them to detect weaknesses and flaws that may go unnoticed by a homogenous internal team, providing a more comprehensive analysis that internal testing alone. By incentivizing the discovery and disclosure of AI model flaws, organizations can build public trust, reduce misuse, 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 safety, security and reliability of AI.

Respectfully Submitted,

Ilona Cohen
Chief Legal and Policy Officer
HackerOne

[8] *Id.*