HPC

HACKING POLICY COUNCIL

Comments to
Request for Information Related to NIST's Assignments
Under the Executive Order Concerning Artificial Intelligence

February 2, 2024

The Hacking Policy Council ("HPC") 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 towards this important proposal.

The HPC is a group of industry experts dedicated to creating a more favorable legal, policy, and business environment for vulnerability management and disclosure, good faith security research, penetration testing, bug bounty programs, and independent repair for security.[2] Many of our members are deeply involved in AI system deployment, testing, and red teaming.

HPC's comments focus on AI testing and red teaming. As AI systems become increasingly common in a variety of environments, including critical and public applications, ensuring the security, safety, and trustworthiness of AI is a major priority. Testing AI for alignment with evaluation metrics is a key safeguard against poor security, discrimination, bias, inaccuracy, and other harmful or undesirable outputs. However, we also emphasize that testing should be only one component of a security and trustworthiness program that includes risk assessment, vulnerability management, incident response plans, and other safeguards.

Below are HPC's recommended areas of focus for NIST as it develops guidance for AI testing and red teaming, as required by the EO.

1. Consistent terminology

NIST should play a role in establishing a common lexicon of AI testing terminology to avoid confusion and enhance collaboration. Greater adoption of a more consistent set of terms and definitions can help ensure that efforts across different disciplines and regions have a shared understanding of the key concepts that govern the testing and usage of AI systems. Consensus on terms also helps auditors and researchers evaluate AI systems based on consistent benchmarks, facilitating easier comparison of testing results and aiding in the development of robust programs to establish trustworthiness and efficacy of AI systems.

[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.

[2] Hacking Policy Council, https://hackingpolicycouncil.org.

We recommend that this language should not be conflated with cybersecurity terms where such terms are not a good fit, such as using "vulnerability" for algorithmic flaws. Distinguishing this language from cybersecurity terms will prevent misinterpretations and help organizational AI evaluation efforts identify and leverage the specific tests that meet their needs.

a. Red Teaming

Red teaming is a term that has taken on a meaning in the AI context that differs from the term's use in other contexts, risking confusion. In the cybersecurity context, traditional definitions of red teaming center on emulating adversarial attacks and exploitation capabilities against an organization's security posture. However, the EO's definition of "AI red-teaming" is not limited to security and does not necessarily require adversarial attacks.The EO defines the term "red-teaming" broadly as "structured testing effort to find flaws and vulnerabilities in an AI system" including "harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system."[3] This definition encompasseses a wide variety of testing, including both adversarial and non-adversarial, to manage both security and non-security risks.[4]

We caution against conflating different types of AI testing as the EO is implemented. While multiple testing methods hold value, their purposes and methodologies are distinct. A lack of clarity regarding testing types risks confusion on the specific type of test required to manage a given risk, especially given that AI demands testing against both security and non-security risks.

We encourage NIST to disambiguate the types of tests covered under "AI red-teaming," and to avoid reference to non-adversarial AI testing methods as "red teaming." It is important to note, however, that this suggestion does not imply that other forms of AI testing outside the scope of "AI red teaming" should be disregarded.

b. Non-security harms

AI systems should undergo multiple testing approaches to address diverse risks and harms that extend beyond the traditional focus on security and safety. This should include testing for non-security harms such as bias, discrimination, fairness, accuracy, and other harmful outputs (such as deepfakes, synthetic child pornography, and toxic content). This is critical to ensure the ethical development and responsible deployment of AI.

However, security and non-security harms should be distinguished in both terminology and testing methodologies. Though at a high level there are similarities in security and non-security risk management functions, such as risk assessments and testing the effectiveness of safeguards, the processes for discovery, validation, and mitigation of non-security flaws differ in practice from analogous processes to address security vulnerabilities. Conflating security and non-security testing metrics and results can interfere with the integrity of the application security pipeline. The organizational personnel tasked with managing these processes for non-security flaws tend to be distinct from the

[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-trustworthydevelopment-and-use-of-artificial-intelligence.

[4] Center for Cybersecurity Policy and Law, AI Hacking in the White House's Executive Order, Nov. 1, 2023, https://www.centerforcybersecuritypolicy.org/insights-and-research/ai-hacking-in-the-white-houses-executive-order.

personnel managing security risks. Security and non-security harms carry different regulatory obligations and liability exposure.[5]

In contrast to security vulnerabilities, there is presently little consensus regarding how to refer to non-security harms or the flaws that can be exploited to enable these harms. We urge NIST to establish more precise technical terms that reflect the different testing methods, mitigations, and legal environments between the two harms. While some practitioners use the term "vulnerabilities" to describe non-security flaws, this risks confusion with security vulnerabilities. The term "vulnerabilities" carries specific meaning in security practices, as well as security regulations and guidance in the United States and internationally, which was not intended to carry over to non-security risks and harms.[6]

Our suggestion is to use "flaws" or "AI algorithmic flaws" as an umbrella term to distinguish non-security harms that are algorithmic in origin from security vulnerabilities. By more clearly delineating between these distinct domains, it becomes easier to implement targeted testing processes and mitigation strategies tailored to the specific challenges posed by each category.

2. External sources: coordinated flaw disclosure, bias bounties

As AI technologies continue to evolve, proactively staying ahead of potential risks is crucial. This involves continuous testing and updates to ensure security, trustworthiness, and fairness in AI applications. While many organizations leverage in-house testing to evaluate AI systems, it will become increasingly important for organizations to be capable of receiving disclosures of both security vulnerabilities and non-security flaws from external sources.

In the context of security, these methods include coordinated vulnerability disclosure and bug bounties. While AI system operators should consider these options for security, AI system operators are also adopting similar processes to receive disclosures regarding non-security harms from unsolicited or external sources. While we are not advocating to require organizations to adopt a specific type of AI testing or monitoring method, we encourage NIST's guidance for AI testing and risk management to encompass methods that facilitate such disclosures from external sources regarding non-security harms: coordinated disclosure and bias bounties.

a. Coordinated flaw disclosure

In security, coordinated vulnerability disclosure programs (CVD) have demonstrated immense value in identifying and mitigating vulnerabilities, minimizing the window of exploitation by malicious actors. In addition to having a vulnerability disclosure process for security, AI system operators should incorporate disclosure and handling processes for non-security algorithmic flaws.

Again, it is helpful to distinguish between organizational processes for disclosure and handling of security vulnerabilities from non-security flaws. Non-security disclosures will require different mitigations than disclosures for security given the distinct nature of non-security flaws, and non-security risks may be managed by a different internal team than security. As noted above, there is not an established term

[5] Hacking Policy Council, AI red teaming – Recommendations for legal clarity and liability protections, Dec. 12, 2023, https://assets-global.website-files.com/62713397a014368302d4ddf5/6579fcd1b821fdc1e507a6d0_Hacking-Policy-Council-statement-on-AI-red-teaming-protections-20231212.pdf.

[6] *Id*.

yet for a non-security coordinated disclosure process. Rather than re-use “vulnerability,” we propose referring to “algorithmic flaws” and “coordinated flaw disclosure” in the context of AI.

b. Bias bounties

Another adversarial method for testing AI systems over time for non-security harms are “bias bounties.” Unlike traditional security-focused bug bounties, which focus on identifying vulnerabilities, bias bounties typically refer to incentive programs aimed at discovering and rectifying bias, discrimination, and harmful output in AI systems. Like bug bounties, bias bounties leverage the broader community to test and disclose flaws to the system operator for mitigation. This approach is especially helpful for identifying issues that may be missed by an in-house testing program.

Just as bug bounties should be backed by a broader organizational vulnerability management process, bias bounty programs are most effective if the AI system operator establishes internal processes dedicated to managing non-security risks that include well-coordinated mechanisms to assess, triage, escalate, mitigate, and communicate about AI algorithmic flaws. Bias bounties represent just one approach among several for testing bias, and they should be integrated into, rather than replace, other testing methods.

3. Diverse testing teams

While automated or AI-driven testing tools play a crucial role in identifying security vulnerabilities and AI algorithmic flaws, human testers are invaluable for validating more complex aspects of AI risk. In the context of AI red teaming, where the objective is to assess and strengthen the system's resilience against potential threats and flaws, leveraging both general human red teamers and subject matter experts is a strategic approach to confirm findings. For example, general testers can identify clearly harmful output, while medical experts may be engaged to identify instances of medical misinformation by AI applications.[7] This collaborative approach, combining the insights of human testers and the collective intelligence of the community, ensures a more thorough and diverse evaluation of AI systems, contributing to enhanced reliability and trustworthiness. The value of varied expertise for comprehensive red teaming indicates the need to leverage the creativity and broader perspective of the community through mechanisms like bias bounties and coordinated disclosure processes to discover flaws that might not be apparent through traditional testing.

4. Testing during development and post-deployment

AI systems should be tested for both security vulnerabilities and non-security flaws during development as well as regularly after deployment. Testing during the development phase helps identify and address security vulnerabilities and can ensure ethical considerations, transparency, and fairness are integrated into the system design before deployment. As with traditional security practices, testing is also appropriate when the AI system is deployed in context, as well as following any significant system changes. Given the evolving nature of AI model training data, attack techniques, mitigations, and features in both security and trustworthiness, it’s important to continue testing post-deployment to address emerging threats and continually enhance the system’s resilience

[7] Slack et. al, A Holistic Approach for Test and Evaluation of Large Language Models, Scale AI, pgs. 5-6, https://static.scale.com/uploads/6019a18f03a4ae003acb1113/test-and-evaluation.pdf (last accessed Jan. 30, 2024).

This is consistent with lessons learned from successful organizational security programs, which often include vulnerability testing during application development, as well as periodic penetration testing and vulnerability scanning of the organization's digital assets post-deployment. A comprehensive testing strategy that spans development and post-deployment allows for adaptive mitigation measures and ensures the sustained trustworthiness of the AI system throughout its lifecycle.

5. Specifying testing phases, components, and objectives

HPC recommends that NIST leverage a common lexicon of AI terms to provide more specificity on the component processes for testing and "red teaming." This includes identifying the type of test that will be performed – whether adversarial or non-adversarial – and delineating its distinct processes and methodologies. The intensity and frequency of testing may depend on the degree of risk associated with the AI system.

Furthermore, identifying the specific goals of the test is essential, including what risks the test seeks to evaluate, how to benchmark severity of risk, and statistical validation metrics for algorithmic flaws. For AI testing to be effective, the testers must have clear objectives regarding what should be flagged as a valid algorithmic flaw, and how to document flagged algorithmic flaws, prior to engaging in the test.[8]

Lastly, a follow-up testing protocol is recommended, involving the re-testing of the model after implementing mitigations or incorporating guardrails established in response to the prior test.

6. Information sharing related to non-security risks

As with security, NIST should encourage AI system operators to establish open communication channels for sharing threats and other information regarding AI algorithmic flaws such as discrimination, bias, or potential harmful outputs. AI information sharing should be informed by guidance and best practices that are already in place, including NIST's AI Risk Management Framework and Cybersecurity Security Framework (CSF). We urge NIST to consider clarifying and adapting its cybersecurity information sharing guidance to encourage information sharing related to AI systems.

Information sharing for both security and algorithmic flaws can foster a comprehensive understanding of risks associated with AI systems, allow for the development of effective solutions to make AI systems more trustworthy, and help organizations be aware of emerging threats and attacks. Additionally, information sharing can contribute to the establishment and adoption of best practices for trustworthy AI across multiple sectors.

* * *

Thank you for your consideration. If we can be of additional assistance, please contact Harley Geiger, coordinator of the Hacking Policy Council, at hgeiger@venable.com.

[8] Andrew Burt, How to Red Team a Gen AI Model, Harvard Business Review, Jan. 4, 2024, https://hbr.org/2024/01/how-to-red-team-a-gen-ai-model.