HACKING POLICY COUNCIL

Comments on the Cybersecurity Artificial Intelligence Community Profile

January 30, 2026

The Hacking Policy Council ("HPC") submits the following comments in response to the National Institute of Standards and Technology's (NIST) initial preliminary draft of NIST IR 8596, Cybersecurity Artificial Intelligence Community Profile ("Cyber AI Profile").[1] We thank NIST's National Cybersecurity Center of Excellence (NCCoE) for the opportunity to contribute to this important initiative.

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.

As NIST develops guidelines for the management of cybersecurity risk related to AI systems and to leverage AI to enhance cybersecurity, HPC's feedback focuses on the critical role of testing and evaluation, and HPC appreciates the inclusion of relevant topics in the initial preliminary draft of the Cyber AI Profile. Specifically, HPC emphasizes the importance of vulnerability management for AI systems and the utility of AI red teaming as AI becomes more integrated into software ecosystems. HPC encourages NIST to maintain the inclusion of these important topics in the final Profile and to expand where possible.

Even as we leverage AI to drive innovation in cybersecurity, we recognize that its evolution introduces novel risks, including in cybersecurity and broader security domains. Given the growing reliance on AI technologies across industries, it is essential to prioritize security and resilience through adequate testing and evaluation. It is imperative that these risks be understood and addressed to safeguard against unintended consequences.

Scope and Terminology Clarification

As NIST develops the Cyber AI Profile (and other, related AI documents) we encourage clear and consistent use of terminology to avoid conflating distinct categories of risk. In particular, distinguishing between cybersecurity risks for AI systems (such as model theft, data poisoning, or prompt injection) and related trustworthiness issues (such as bias, hallucinations, or

[1] Megas K, Cuthill B, Snyder JN, Patrick B, Khemani I, Dotter M, Garris M, Zarei M, Schiro N (2025) Cybersecurity AI Profile: NIST Community Profile. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Internal Report (IR) NIST IR 8596 iprd. https://doi.org/10.6028/NIST.8596.iprd
[2] Hacking Policy Council, https://hackingpolicycouncil.org.

undesirable outputs that may pose cybersecurity risks). While both these categories merit robust risk management, they differ in how they are discovered, validated, mitigated, and disclosed. Organizations integrating AI risk management into existing cybersecurity programs would benefit from clear discussion of both categories, and how the Profile applies (or does not). This clarity will improve interoperability with existing NIST frameworks, international standards, and vulnerability disclosure practices, while supporting effective operational implementation.

Vulnerability Management for Artificial Intelligence

Vulnerability management is a fundamental and necessary aspect of cybersecurity. In previous comments, HPC has repeatedly advocated that vulnerability disclosure and handling processes should be explicitly included in NIST guidance and standards.

Vulnerability Disclosure Policies (VDP) have demonstrated immense value in identifying and mitigating vulnerabilities, minimizing the window of exploitation by malicious actors. While many organizations leverage in-house testing to evaluate 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.

With the introduction of novel risks posed by the adoption of AI and the need to manage such risks, the value that VDPs pose for strengthening security has only increased. A robust vulnerability disclosure process ensures that systems that have incorporated AI are continuously improved, and increases the likelihood that potential risks are mitigated. This proactive approach helps maintain the integrity, reliability, and trustworthiness of AI systems, safeguarding users and stakeholders from unintended harm or misuse.

The Cyber AI Profile would be strengthened by explicitly acknowledging that organizations may need tailored communication channels and handling processes for different classes of AI-related risks, while maintaining alignment with established coordinated vulnerability disclosure practices for security vulnerabilities. The Cyber AI Profile should recognize that AI systems may require multiple, differentiated disclosure and handling processes. While traditional vulnerability disclosure policies are well-suited for cybersecurity vulnerabilities, AI systems may also present non-security flaws or misuse pathways that warrant separate intake, triage, and response mechanisms based on cybersecurity characteristics.

For example, disclosures related to model behavior, misuse potential, or algorithmic flaws may require evaluation by different internal teams, distinct mitigation strategies, and different timelines than cybersecurity vulnerabilities. Treating all AI-related issues as a single category risks misrouting reports, delaying mitigation, or prompting inappropriate disclosure decisions.

HPC appreciates NIST's incorporation of vulnerability disclosure handling processes in the initial preliminary draft of the Cyber AI Profile. Within the IDENTIFY section, the Profile recommends that "processes for receiving, analyzing, and responding to vulnerability disclosures are established." For the "thwart" focus area, the Profile establishes this is a high priority, noting that

“vulnerability management, disclosures, and response will need higher prioritization due to speed and efficiency of AI-enabled attacks to exploit vulnerabilities at scale.” The Profile also references vulnerability management in the GOVERN portion, recommending organizations “create AI-specific communication channels for sharing information about threat vectors and/or defense capabilities.”

HPC appreciates the Profile’s recognition of the importance of vulnerability management. Given its necessity, HPC encourages NIST to maintain these references in future versions of the Cyber AI Profile, and where possible further incorporate them throughout the document.

Leverage AI Red Teaming

With an increase in both threats to AI and threats leveraging AI, conducting thorough testing and evaluation is crucial to strengthen trustworthiness and resilience. HPC believes that this should include red teaming exercises to rigorously test AI systems, whether using in-house teams or third-party service providers. Such practices help identify potential vulnerabilities and weaknesses, ensuring that AI models are robust, secure, and reliable.

While organizations may be familiar with red teaming to test software for security, “AI red teaming” typically has a broader scope in relation to cybersecurity, to include flaws and vulnerabilities or other harmful or undesirable outputs. By identifying and disclosing misalignment with potential cybersecurity implications in AI systems so they can be corrected, AI red teaming is a beneficial practice to help ensure the security, safety, and trustworthiness of AI.

HPC also encourages organizations to include clear protections for individuals and teams conducting authorized or good faith AI red teaming and testing activities. These protections may include safe harbor language, clear authorization boundaries, and commitments to avoid punitive or retaliatory actions against researchers acting in good faith and within defined parameters. Without such assurances, organizations may discourage rigorous testing or limit red teaming to narrow internal exercises, reducing its effectiveness.

HPC also urges caution in how AI red teaming results are documented and shared. Detailed findings from red team exercises may expose unmitigated weaknesses, testing methodologies, or attack pathways that could be misused if disclosed broadly or prematurely.

Consistent with established cybersecurity best practices, the Cyber AI Profile should emphasize that documentation of AI testing results should be proportionate and risk-based. Summary-level reporting may be appropriate for governance, oversight, and continuous improvement purposes, while detailed technical findings should be tightly controlled and shared only on a need-to-know basis.

Explicitly recognizing the sensitivity of AI testing outputs will help organizations balance transparency with security, and reduce the risk that well-intentioned testing efforts inadvertently create new attack vectors.

HPC appreciates NIST's inclusion of AI red teaming in the initial preliminary draft of the Cyber AI Profile. In the IDENTIFY section, the draft recommends for organizations that "improvements are identified from AI Red Teaming exercises and through coordinated testing with suppliers of AI models and data to harden the supply chain." HPC appreciates these references to AI red teaming and encourages NIST to maintain them in the final Profile.

Bug Bounties for AI

As AI technologies continue to evolve and adoption increases, proactively staying ahead of potential risks is crucial for organizations. HPC supports the use of bounties to incentivize researchers to identify and disclose AI model vulnerabilities and flaws. In the final Cyber AI Profile, HPC encourages NIST to specifically incorporate recommendations around bug bounty programs.

By expanding bounties beyond traditional cybersecurity risks to include non-security issues and misuse, organizations can foster proactive independent research that makes AI systems safer and more trustworthy. This expansion is particularly important because AI systems are susceptible to a range of risks that traditional software may not encounter, such as data poisoning and the manipulation of AI outputs.

In implementing bug bounties, specifically with a broader scope, organizations can tap into the expertise of the global security research community to identify and mitigate misuse risks early, before they can be exploited in real-world scenarios. This proactive approach not only enhances the security of these AI systems but builds trust in their deployment.

HPC encourages NIST to frame bug bounty programs as one indicator of a mature and proactive AI risk management posture, rather than as a baseline expectation for all organizations. Effective bounty programs can vary significantly in scope, structure, and resource requirements, and flexibility is essential to ensure broad adoption across organizations of different sizes and risk profiles.

Lifecycle and Change Management Considerations

HPC encourages NIST to emphasize that cybersecurity risk management for AI systems must extend beyond initial development and deployment. AI systems frequently evolve through model updates, retraining, fine-tuning, data refreshes, and integration into new operational contexts, each of which may introduce new risks or invalidate prior testing assumptions.

The Cyber AI Profile should explicitly recognize the importance of post-deployment monitoring, periodic reassessment, and re-testing following material system changes. Integrating AI risk

management into existing vulnerability management, patching, and incident response processes will help organizations maintain resilience over time and avoid a “one-and-done” approach to AI security.

Framework Alignment

HPC encourages NIST to continue prioritizing harmonization between the Cyber AI Profile and existing cybersecurity frameworks, including the NIST AI Risk Management Framework, Secure Software Development Framework, and relevant international standards.

Organizations benefit when guidance is additive and interoperable rather than duplicative or conflicting. Clear alignment will reduce implementation friction, enable reciprocity across frameworks, and help ensure that investments in existing security programs can be leveraged effectively as AI systems are adopted.

By positioning the Cyber AI Profile as an extension of established risk management practices, NIST can accelerate adoption while avoiding unnecessary complexity or compliance fatigue.

* * *

HPC appreciates the opportunity to provide these comments and welcomes further conversation. For additional information, please contact Heather West, Venable LLP at HEWest@Venable.com.