Alicia Chambers
NIST Executive Secretariat
National Institute of Standards and Technology
100 Bureau Drive, Stop 8900
Gaithersburg, MD 20899

RE: Palo Alto Networks' Comments in Response to NIST Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence, NIST–2023–0009

Introduction

Palo Alto Networks appreciates the opportunity to provide comments in response to the Request for Information (RFI) related to the National Institute of Standards and Technology (NIST) Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (NIST RFI).

Palo Alto Networks is the global cybersecurity leader. We were founded in 2005 and have since become the world's largest cybersecurity company – protecting businesses and government agencies across more than 150 countries. We support 95 of the Fortune 100, critical infrastructure operators of all shapes and sizes, the U.S. federal government, universities and other educational institutions, and a wide range of state and local partners.

Our company firmly believes the risky outcome for society would be to *not* meaningfully leverage AI for cyber defense purposes, and Palo Alto Networks is aggressively innovating and investing to ensure we can continue delivering superior security outcomes. Our product suite – which spans network security, cloud security, endpoint security, and security operations center (SOC) automation – has successfully leveraged AI and machine learning (ML) for many years to stay a step ahead of attackers.

Every day, we analyze over 750 million new and unique security objects. Additionally, every day we detect over 1.5 million unique attacks that are novel or previously unseen. This process of continuous discovery and analysis allows threat detection to stay ahead of the adversary. This real-time awareness of the threat landscape allows our company to block over 8.6 billion attacks each day. This would simply not be possible without AI.

Cyber adversaries are already leveraging AI to advance their tradecraft and will continue to do so going forward, which makes AI's role in defensive cybersecurity even more vital. AI supercharges cyber defenses and helps defenders anticipate, track, and thwart cyber attacks to a degree never seen before. Even in the case of generative AI, LLMs can bring considerable

advantages to cyber defense. For example, LLMs make it easy to process large amounts of information to better identify threats and vulnerabilities in a sea of data. They can provide tremendous efficiency, intelligence, and scalability for managing vulnerabilities, preventing attacks, handling alerts, and responding to incidents.

We encourage all entities to embrace the adoption of AI for this critical use case. As NIST considers developing guidelines, standards, and best practices for AI safety and security, Palo Alto Networks urges a risk-based, use-case specific, and stakeholder-involved approach that prioritizes the prevention of unintended consequences for beneficial use cases such as defensive cybersecurity. A number of flexible frameworks have been advanced that could help organizations identify and contextualize features of AI models and associated risks for particular use cases. For example, useful factors for consideration include aspects like the predominant nature of the data processed by the system; whether an AI system is facilitating human decision making; or if the AI system is supporting consequential decisions with human impact (e.g. employment, creditworthiness, etc.) versus an AI system that enables network processes or makes them more efficient, among others.

NIST and other agencies tasked with implementation of the executive order must carefully consider the alignment of common principles and standards that will enable a targeted and flexible governance approach across the global AI regulatory landscape. Such an approach could help minimize conflicting requirements. It could also promote innovation, research, and development.

Given our role as a leading cybersecurity provider, Palo Alto Networks offers the below information for consideration in response to NIST's RFI on AI Security. Specifically, we provide information on two critical areas addressed in Section of 1(a)(1) of the RFI: information on model validation and verification, and AI red-teaming; and helpful controls for various risk management purposes.

Our comments are categorized into two principal sections, each addressing areas where Palo Alto Networks has extensive experience and expertise in both using AI for cyber defense purposes and testing AI models and systems for vulnerabilities or adversarial exploitation. The first section outlines tools and techniques we employ to test and validate generative AI models against adversarial use, and to conduct red-team evaluations. The second section contains best practices and controls used by our risk management teams to identify and mitigate AI model risks prior to deploying those models in a production environment.

AI Model Validation/Verification and Red-Team Metrics

Given our deep experience in the cybersecurity field, Palo Alto Networks has extensive expertise in developing and conducting model validation, verification, and red-team exercises, as well as AI-augmented and AI-automated validation, verification, and red-teaming of

generative AI systems. Through these efforts, we test both the security of networks and information systems and the ability of organizations to maintain the confidentiality, integrity, and availability of those critical networks and systems. As applied to generative AI systems, organizations can employ many AI-enhanced testing best practices used by cyber practitioners but must adjust some of those practices to address unique aspects and capabilities of generative AI systems.

Isolated Testing and Evaluation Environments

First and foremost, because using generative AI to adversarially test systems requires running and generating attacks, it is essential to ensure that AI-enhanced red-team assessments that include adversarial AI are done in a safe environment. This environment must be one that is disconnected from production environments and other areas where organizations may be processing proprietary or sensitive data, like customer information or other forms of controlled data. This is especially important as AI can have advanced capabilities to discover and exploit weaknesses, mask indicators of compromise, and evolve characteristics to avoid detection. We strongly recommend segmented networks that are clearly identified and marked by testing teams as “unsafe” and physically isolated to prevent unauthorized access to any confidential or sensitive data.

To that end, a recommended best practice is the use of a “dirty lab” or “safe testing” environment with strong security guardrails to perform AI attack analysis and observe malicious traffic resulting from the use or operation of the relevant AI systems. The “dirty lab” should allow users to run malicious and untrusted software to prevent adverse impacts on corporate networks or environments where organizations store proprietary or sensitive information. It is critical that this testing space is completely separated from environments with customer data or sensitive company information. This segmented network environment should also be monitored for any communication attempts made to external systems, which is indicative of command and control or data exfiltration behaviors.

If malicious AI signatures and behaviors are identified, deeper analysis of those signatures and behaviors on other relevant endpoints is critical to better understand the AI attack. If localized exploration is required or desired, perhaps for edge AI intended to run on endpoints or for container-based AI systems, we recommend using a localized virtual machine (VM), with all networking disabled, to evaluate and analyze the malicious content. Additional measures that should be employed to protect critical networks and environments include, but are not limited to, ensuring the host operating system is incompatible with the operating system of the VM to prevent spill-over issues; utilizing endpoint protection tools on the host that monitors malicious activity; and clearly labeling and isolating malicious signatures to both categorize threat information and operationalize information across defensive systems.

Top-level isolation of these testing environments will help ensure they are self-contained so adversarial AI does not compromise or impact other critical systems or critical data that organizations use as part of their business or production environments, and does not reach out onto the external internet.

Best Practices for AI-Augmented Evaluations

For generative AI systems, AI-augmented evaluations should cover relevant threats or unique aspects that open systems to adversarial use or exploitation. As such, teams should ensure their analysis includes a variety of aberrations such as: truncated responses of the generative AI system, or query responses that stop in the middle of a sentence; deviations in conversational or human-like chatter; incorrect tool requests and responses, like instances where the generative AI tool passes non-responsive or incorrect responses (e.g. passing multiple vulnerability signatures in response to a request to pass a single signature); instances where the generative AI system incorrectly gives-up or decides it cannot action a given prompt; hallucination, which, in the cybersecurity context, would include where a generative AI system is being used to forecast how a vulnerability can be weaponized, and the system responds with information on a hallucinated vulnerability that is unrelated to the prompt.

Process-wise, it is preferred for these evaluations to operate with as much automation as possible to ensure they are able to run continuously as the generative AI system learns and responds to testing and evaluation. As part of this work, a best practice is to leverage mock network calls in order to improve performance and increase reliability of tests. Human evaluation must be used to spot check the machine evaluation, to ensure it is correctly converging but also to uncover novel cases or issues. Testing teams should be mindful, however, that these human-led processes can be slower and less repeatable than automated processes, but is nonetheless important for AI system validation.

Red Team Metrics to Evaluate Model Resilience and Inform Risk Management

Red-team testing processes should be supported by clear and understandable metrics to measure multiple attributes. At the outset, we have found that an effective initial metric for red-teams to assess is the AI system's fragility against attacks by quantifying the red-teams success in simulating an attack. This can generally be framed as measuring the attacker's cost of exploiting an AI system against their potential gains in a successful attack. Specific to generative AI systems and Large Language Models (LLMs), total costs could entail measuring the attacker's labor, financial, and time resources spent to access a particular model and the associated network and compute costs. For open source models, the attacker's costs will generally cover any compute or storage requirements associated with use of the AI system.

The attack-focused, cost-benefit metric should then be factored against the relative cost to the business of having their AI system compromised by adversarial actors. This can include

resources and investment required to fix the root cause of the exploited system or vulnerability, and we have seen this business-cost metric influence the persistence of attacks. In other words, attackers are assessing how costly it would be for businesses to address and fix certain exploitations, and assessing how likely it would be for victims to address those issues when making decisions about using a particular attack vector. For red teams assessing generative AI systems and models, this can present a challenge because the team will need to continue exploiting the system or exfiltrating data until they have fully assessed the root cause of the vulnerability or attack vector. As a result, red teams should be mindful that this continued exploitation and/or poisoning of model data can change model behavior, which could impact the utility and performance of the model. When done in an appropriately segmented and isolated environment, red teams' identification of these root causes is essential to evaluating the likelihood and persistence of attacks, which gives the AI model provider the ability to identify and prioritize risk.

Other accompanying evaluation metrics should include: accuracy of the responses and answer relevance, specifically whether the tool understood the prompt and correctly answered the query; no-answer rate, including both the tool's inability to generate a response to a prompt, but also to validate where the tool will state it cannot respond to a prompt rather than developing a hallucinating or non-relevant response; the incorrect response rate, to include the number and frequency of false positives; and answer completeness, or measuring where the tool responded to all aspects of the prompt. Importantly, and to gauge the propensity of a generative AI tool to be exploited for adversarial use, the prompts for the AI evaluation set should be curated to include actual and expected user input, intentionally malicious input, and off-topic questions and prompts. Prospectively, we are exploring where generative tools can be used by red-teams to develop evaluation prompts to increase metrics and coverage of the evaluation tool.

Helpful Controls for AI Risk Management Processes

AI risk management processes should ensure that the use and operation of a given tool concurrently protects an organization's data while ensuring both the tool and critical AI assets are protected against threats and vulnerabilities. Risk management processes should focus on securing each stage of the AI system lifecycle in a production setting, encompassing data collection, data processing, model training, model validation, and model deployment. Additionally, risk assessment processes should address additional aspects like the AI supply chain and the establishment of controls and policies for backup, recovery, and contingency planning related to the given AI tool.

Specific to AI development, we believe that the initial effective step involves identifying and classifying the most critical data, assets, or systems within an organization. These components, if breached, harmed, or lost, can pose significant threats to the organization's operations, reputation, or competitive standing. This list can encompass: data sets utilized for training, refining, and evaluating AI models; prompt templates, which are the result of

prompt-engineering; few-shot examples (e.g. prompt techniques that facilitate in-context learning for Large Language Models (LLMs) that can improve performance and response accuracy); scripts, such as Python or JavaScript, utilized for model training; fine-tuned AI models; vector databases; storage environments employed for the organization's enterprise search capabilities; and the AI application user's chat history and feedback collected for improving the model.

Suggested Data Controls

As a best practice, we encourage NIST to facilitate risk management processes that specifically address the following categories of controls to evaluate and secure an organization's AI tools and capabilities.

Training data protection: The training data utilized in AI systems should be stored securely to prevent data leakage from the AI environment. Teams should validate that the training data is encrypted at rest both in the all primary and intermediate storage environments of the data processing pipeline. Validation should include ensuring the training data can be transmitted only using encrypted and secure communication channels. Lastly, continuous monitoring of the training environment must be enforced to detect suspicious activity like unknown or unplanned data downloads or the transmission of data to unapproved devices.

Access control: Organizations should ensure they implement and maintain access control processes and best practices on all training data throughout the data processing pipeline. We have seen there can be a propensity for organizations to lose access control when data is transferred from its original source into the AI data processing pipeline. Oftentimes, this can be attributed to inconsistencies in access control measures between data processing environments, or not ensuring initial measures are enforced across downstream processing environments.

Particular attention should be paid to data environments where organizations process sensitive data to ensure such processing is occurring in separate and secure environments. If multiple datasets are being processed in the same AI environment, organizations should consider utilizing several service accounts. Each account should have controlled access to a specific subset of the training data, with granular permissions based on the sensitivity of the data so that even if a single account is compromised, it would not compromise all of the data in the secure environment. Third party resource providers, contractors, or other external parties should not have access to an organization’s training/test data assets without contracts in place that fully articulate how, where, and when parties can access and use training data. Auditing access to training data, an essential control to ensure and validate the confidentiality of the data set, should be done at each stage in the data processing.

This access to training environments should be limited to AI research teams, and organizations should ensure those teams do not have access to relevant production environments. Automation in the production of AI systems should be maximized, with resource access limited to service accounts, thus reducing the requirement for human intervention.

Training data poisoning protections: Training data collected from untrusted sources for model training could contain sensitive personal data, proprietary data, inaccurate data, or other data that could affect the performance of a model or present compliance risks to the organization. As such, risk management processes should ensure that data used for training is collected from trusted sources and goes through the organization's AI governance process prior to use in an AI system. To that end, and where applicable, data pulled from the internet or other untrusted sources should go through a filtering process to identify malicious data that would negatively impact model characteristics. Open data sources should go through the organization's governance process to ensure training data sets are protected against poisoning and backdoor attacks. Similarly, inference data provided by external users used to retrain a model should not be implicitly trusted and should be treated as new data. Throughout, all training data sources should be documented, logged, and made available with the data set to facilitate auditing. Any auditing should document the responsible data set owner and how the owner adheres to documented data protection policies.

Sensitive data protection: Additional protections and controls should be included in the risk management process where AI tools would be using sensitive data, like certain categories of personal data or proprietary commercial data essential to the organization's operations. For instance, only data pertinent to the AI model's functionality should be moved from the secure production settings to the development environment, post adherence to the organization's AI governance procedures. Before transferring to the development environment, some data types might need pre-processing in a secure location, during which any sensitive information, such as personal details, should be anonymized. Finally, individuals engaged in the utilization or development of AI models with sensitive data should undergo specialized training on data policies, safeguards, and documentation procedures.

Data integrity: Organizations must ensure that any alterations to data sets used for training or fine-tuning the model during the AI lifecycle are auditable and trackable to maintain control over these environments. This protects the integrity and reliability of the AI model and aids in tracking the model's performance, identifying potential issues, and ensuring that the model is functioning as intended. Collected data sets must be accompanied with some additional information known as metadata that can provide users with necessary information about the dataset. Metadata should have a unique identifier, version of the data set, relevant description on the data and a digital fingerprint of that data that enables users to validate the integrity of the data and any other useful labels. Datasets should be uniquely identified so risk management processes can detect unauthorized changes to an approved data set, and where appropriate, initiate a comprehensive review.

Suggested Controls for AI Models Prior to Public Consumption

Prior to publicly offering AI models to customers or before including models in production, Palo Alto Networks suggests implementing the following controls as a best practice for risk management teams.

Model deployment: Risk management processes help ensure models are trusted both at the deployment stage and throughout the AI system lifecycle. In order to facilitate this compliance, organizations should implement strong access controls by allowing only authorized individuals, after going through the standard approval process with the appropriate AI governance team, permission to deploy the model into the production environment. AI systems should automate the build and deployment of the models. Only restricted and pre-approved accounts should be allowed to access these build systems. Organizations should ensure AI models are deployed on trusted platforms that comply with industry-standard security practices, and only with vendors that have been vetted and approved with risk management professionals. To prevent stealing of the AI model's attributes once deployed, high-precision confidence values should be randomized or masked in outputs to prevent threat actors from copying or training models that function similar to the deployed model.

Model access control: Without proper access control on models, unauthorized entities can duplicate, simulate, or reconstruct a machine learning model, leading to intellectual property theft, inferring training data based on the model's outputs, or misuse of the model for malicious purposes. For AI models accessed through an application protocol interface (API), risk management controls should be implemented to ensure access is managed through identity and access management (IAM) policies. Additional controls that limit the number of calls (rate limiting) must be enforced to prevent abuse of the AI model. To that end, access logs should be regularly audited to enable risk management teams to detect unauthorized access to the model or any of the associated data.

Model versioning: A versioning and tracking model is critical for risk management teams to investigate, and where appropriate, reproduce an incident using a particular AI model. New versions should be assigned and documented every time a model is trained, and old versions should be deprecated in the production environment. Last, model versions should employ qualifiers to differentiate pre-production/development models from the model that is offered in the production environment. This practice can also reduce errors and vulnerabilities when a model changes unexpectedly, which can change the characteristics of the AI system because its AI model components changed without being retested. Proper communication and change management with users of a model when a version of a production model will be changing is essential to avoid unpredictable and potentially insecure behaviors from arising without the downstream user's knowledge.

Privacy and access control: To prevent threat actors from gaining access via an AI model they otherwise would not have access to, organizations must ensure that the users are properly authenticated in the preceding application layers before the call signal can reach the AI model itself. To prevent attackers from accessing customer data, user data, other model configuration details, or any application data, all responses produced by the AI model should undergo a thorough validation process before being used in subsequent layers of application. All instructions generated by the generative AI application, especially those that are asked to retrieve customer or other personal data, should be executed through a particular user's authentication token, using the exact same process as the requesting user. For example, many

generative AI applications translate users' questions into API calls or structured language queries (SQL) to fetch the data that the user is requesting. It is critical that these instructions are executed with the permissions of the requesting user or user role rather than a high-level application user who has more access rights. Otherwise, the attackers can craft creative prompts and gain improper access to data. AI risk management controls should include components that monitor and evaluate data access and user activity.

These practices also enable risk management teams to audit activity logs and track user access. The application must also contain well-defined and secure error handling mechanisms to prevent the leakage of sensitive information. As discussed above, an additional protection that is available is for the AI model to use anonymized or de-identified personal data to mitigate the harm of unauthorized data access.

Access controls for AI chat history: An often overlooked attack vector available in certain AI model applications is the chat history, which maintains an ongoing log of prompts and queries used by an organization and its AI model users. Chat history can be used both in inference and training. Therefore, it can be an attack vector if compromised. This data may also be particularly sensitive as it is provided by users and/or customers and may contain personally identifiable information. Risk management controls should implement robust data access controls and encryption measures that help ensure only authorized individuals can access this chat history. Such controls can be implemented as role-based access controls, where organizations can limit the number of individuals that can access and view chat history. Role-based access controls can facilitate auditing processes to identify who within a given team is accessing chat histories, and give appropriate team members the ability to sanitize and/or de-identify sensitive personal data contained in chat history. Additionally, access controls to chat history can give legal teams the ability to review logs before using this data for subsequent training or other internal learning processes.

Conclusion

Given our extensive, years-long experience with AI technology, Palo Alto Networks looks forward to being an active participant in NIST's efforts to implement the Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. We would be happy to discuss our submission, the defensive cyber use case, and nuances of AI and cybersecurity in greater detail. For more information, please contact: Sam Kaplan, Assistant General Counsel, at skaplan@paloaltonetworks.com.