Openpolicy

Feb. 2, 2024

To: The National Institute of Standards and Technology (NIST), via ai-inquiries@nist.gov, *Atten: elham.tabassi@nist.gov.*

OpenPolicy comments on

Request for Information (RFI) Related to NIST's Assignments under Sections 4.1, 4.5 and 11 of the Executive order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Overview

OpenPolicy appreciates the opportunity to provide feedback on the NIST RFI under the AI EO. OpenPolicy is a technology company seeking to democratize the ability of innovative companies of all sizes to engage with policymakers and provide feedback on relevant policy deliverables. OpenPolicy is further engaged in emphasizing the need to use AI, innovation, and technology to foster open policymaking and broader use of technology to streamline compliance and governance automation. We anticipate active engagement with NIST throughout the evolution of this RFI, and are available to provide further feedback as needed. We are committed to supporting the implementation of the Executive Order and actively engaging with NIST and other implementing agencies, alongside our engagement with the NISA AI Safety Institute Consortium. Indeed, our commitment to work with NIST is documented in the White House AI Executive Order ("AI EO") page itself.[1]

We believe that the open and collaborative nature of the policymaking dialogue is essential to further the implementation of the AI EO, which can be significantly contributed by the participation of innovative companies and, specifically, "startup" companies that develop cutting-edge AI security and safety solutions and AI-enabled solutions. Indeed, many, if not most, of the technologies used to support the requirements of the AI EO and relevant OMB and the underlying NIST guidelines referred to, evaluate the measurements, testing, and audibility of AI, data and security posture, and facilitate the secure adoption of AI and sharing of data, more broadly, are developed by such innovative companies and startups – ***these are the communities OpenPolicy collaborates with***.

[1] White House, "What are they Saying", https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/31/what-they-are-saying-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/ (Oct. 31, 2023).

Openpolicy

While are agree ISO/IEC foundational SC42 body of standards is key for NIST work on this RMF, we note the limited stakeholder engagement from innovative companies in standards development bodies - OpenPolicy is excited to collaborate with NIST to explore the use of AI to cultivate participation in both policy development and standards, including by adopting ML-readable formats to published controls and documents, and facilitating mappings.

OpenPolicy applauds NIST for its leadership in implementing the AI EO requirements. We are especially satisfied with the emphasis on AI security, AI red teaming, and governance and automation in these areas, as well as the need for specific, actionable responses from stakeholders to shape initiatives for evaluating and auditing AI systems.

Today, we believe that one of the most critical gaps is fostering enhanced measurability of AI risk, and AI security and safety outcomes, a better understanding of the holistic risk of AI, and cultivating a ***technology-scaled/centric*** approach to AI risk management and controls alignment/compliance. Moreover, it is key to acknowledge the interconnected nature of AI ***with data***, and the need to facilitate secure and safe data sharing and posture management, as well as ensuring alignment with data security measures that are essential for data set protection, such as measures to protect from ***cryptography risks***.

While we provide detailed comments below, we would like to propose the paradigm of risk management under the NIST CSF (2.0)) and NIST RMF needs to be further updated to address the holistic risk presented by AI and the full life cycle of the AI system, as part of this inquiry.

While the current focus of the RFI seems to address AI risk while focusing on AI, as a technology paradigm - **such a framework is insufficient to encompass the threats elevated by AI** (for example, due to the combination of AI threats with broader datasets, software, IoT threats or bot-scaled attacks), or the underlined context of risk AI operates in (in relationship to certain, data or model, or in combination with a bot attack/bot interaction). NIST should consider a broader risk framework that takes into account these threats and the extended threat surface, as well as the amplification of AI on the risk posed across all technology paradigms.

Openpolicy

Risk measurement and controls must be scaled by technology and automation to scale compliance and governance

OpenPolicy encourages NIST to explore additional opportunities to leverage technology and automation tools to ensure adoptres' practices align with advanced responsible and secure AI posture - as they seek to adopt controls. Put simply, it is not sufficient to simply lay down controls for AI security and risk measurement - if we do not have a scaled approach to ensure such controls are adopted. A risk-based approach can only be scaled, given the nature of AI threats constantly evolving, by leveraging solutions to dynamically assess the risk of the system/model vis-a-vis the controls used. Governance based on documentation alone, has proven is not sufficient. The approach for the control measurement, under the AI EO and RFI, consistent with the direction of NIST CSF 2.0 - ***needs to include a key pillar of governance which is supported by automated means.***

Such an approach can benefit from tools like OSCAL for ensuring control mappings as more controls are added, and the ongoing implementation of the Cyber Executive Order and other government initiatives.

More robust measurements scaled by run-time monitoring of the risk and controls applied, alongside third-party tooling and testing, are needed to scale compliance, and compliance measurements are needed to ensure agency preparation, alongside, of course, ample budget and workforce.

Further, these measures could also foster more accurate and scaled techniques for model validation, human rights impact assessments, and controls for public application presentation which are crucial aspects and practices of risk management strategies and secure development and deployment of AI, and considered in the AI OMB memo. ***For further context, we attach herewith as an appendix, our OMB memo comments as well.***

Risk management, governance, and assessment methods that can adapt with controls released can enable broader monitoring agency implementation.

Automation can further support the complex goal of aligning the implementation of the AI EO, with the Cyber Executive Order with the ongoing implementation and upcoming revisions of FedRamp FISMA, FISMA 2024 priorities related to IoT and OT protections (where AI can be leveraged), and other NIST documents such as CSF 2.0. A further focus on leveraging automation and measurable compliance (e.g., schemes like OSCAL) can advance risk governance

in this respect and increase the ability of NIST to scale agency measurements, as agencies and auditors can collaborate with solutions providers to automate control

measurements.

Our ecosystem of innovative companies specializing in AI security and risk mitigation, is eager to work with the administration to explore how technology and automation can best support the deployment of AI, security, and privacy-related controls and their effective enforcement and measurement. OpenPolicy is also eager to work with NIST to explore how control parsing can further better public-policy engagement on specific controls.

Alignment with other ongoing initiatives and languages

The implementation and development of the RFI deliveries should align with upcoming revisions to other NIST guidelines and controls and recent agency (e.g, CISA "Secure by Design") published best practices and foster further coherence between the Cyber EO (14028), AI EO, and related efforts (such as CMMC) controls' deployment and standards by NIST, as it relates to guidelines development and enforcement efforts. Notably, ***the CSF 2.0 revisions should already take into account proposals steaming from this RMF,*** given the relationship between SSDF, and CSF 2.0.

Furthermore, as part of the future and forthcoming revision of the White House ***Zero-Trust Framework***, the threats and protections of AI must be considered. Beyond this effort, the entire body of cybersecurity controls that apply to agencies needs to be considered in this context, or else the critical work done under this RFI, we fear, may not achieve its full goals.

By way of example, as NIST undergoes implementation of the AI EO, several key related documents that outline controls related to the Memo are being updated to comport with the threat landscape. Notably, the Cyber EO (14028) attestation form, currently under development by OMB, refers to NIST SSDF (SP 800-218), which is expected to be revised under the AI EO. NIST SP 800-171 is also undergoing revisions. Failing to consider the AI RFI effort in the context of these already ongoing efforts risks creating a piecemeal approach that will face challenges in addressing the current threat landscape simply due to separation between cybersecurity and AI governance related controls, agency budgets, risk frameworks and audit mechanisms.

It is key that the RFI consider the need for coherence between the implementation of relevant cyber requirements and newly introduced AI cybersecurity requirements and controls under the ongoing implementation of the Cyber EO (14028) while we await final details of implementation related to AI EO, to ensure agencies and

contractors are implementing the relevant controls, and not further cultivate adherence to past methods, that may expose federal systems to threats.

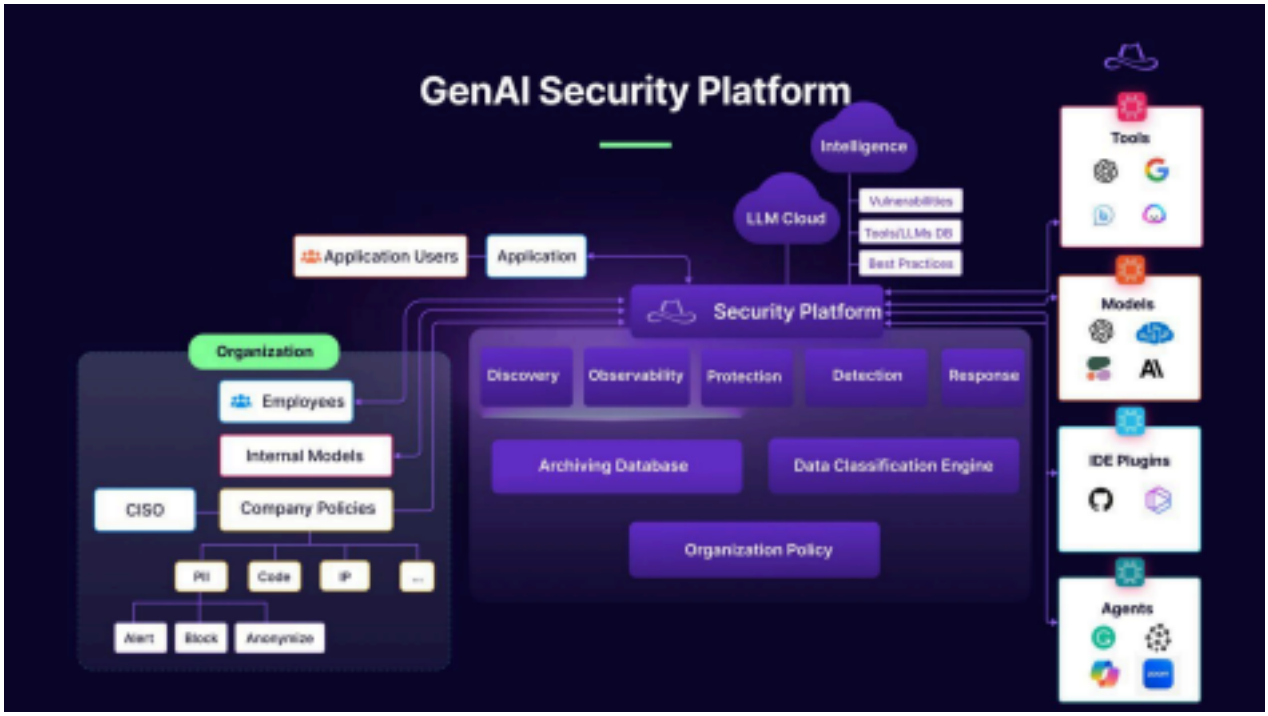
As an additional example, further consideration should be given to whether the Zero Trust Framework architecture by NIST needs revisions to align further with the newly developed AI requirements.

Standards, implementation and proposed controls

Implementing real-time detection for immediate threat notification and response, and adoption a full life cycle risk management approach to AI threats and LLM -

In addressing the risks of synthetic content, under "*Reducing the Risk of Synthetic Content" and broader AI and LLM threats,* it is crucial to establish robust security governance, including data classification and access control, particularly for LLMs.

Furthermore an holistic approach to addressing LLM threats needs to consider the full model and AI cycle (see below scheme). It should further include measures implementing real-time detection and alerts, monitoring LLM data flows for unauthorized disclosures, and enforcing comprehensive protection policies - including monitoring the risk posture based on event monitoring, analysis, context on user behavior and anomaly and threat detection.



Openpolicy

[***LLM security lifecycle illustration, source: Lasso Security – note: icons are just for illustration***]

This approach should encompass detailed data transmission tracking and stringent security measures, providing a holistic defense against LLM threats and supporting secure and trustworthy AI development and deployment lifecycle at different levels of the AI system and in different modes of model deployment as the section address. It should also include governance and auditing of which data sets are unstructured content and data is being used by am AI systems that can build on SP 800-171 controls.

Red Teaming and Testing Teams Integration - In the quest to boost AI system dependability, integrating automated tools with the nuanced understanding of human testers in AI red teaming efforts is crucial. This approach includes enlisting both general red teamers for broad flaw detection and specialists for pinpointing nuanced issues such as misinformation and deepfakes. By fostering a collaborative environment that draws on the community's collective wisdom, thorough evaluation, varied skill sets, and encouraging practices like bias bounties and vulnerability disclosures, NIST can uncover otherwise unnoticed flaws.

Further, NIST can consider to develop a unified ***approach for AI testing,*** building on approaches such as MITRE-ATLAS, clearly distinguishing between cybersecurity and

AI-specific terms to support practical evaluation of AI systems. This includes developing and differentiating testing methodologies for security and “non-security harms”, such as bias and discrimination, and establishing terminology for algorithmic flaws. As HPC submission clarifies, this includes distinct definitions and testing methodologies for traditional cybersecurity threats versus AI algorithmic flaws, ensuring comprehensive evaluation of AI systems. Establishing clear standards and language for "AI red-teaming" can further distinguish between adversarial and non-adversarial testing approaches, enhancing the efficacy and trustworthiness of AI technologies.

AI continuous secure deployment- Addressing security vulnerabilities and ethical considerations upfront and continuing to monitor for emerging threats are crucial measurements to bolster AI system safe and secure deployment, thereby emphasizing the need to conduct testing for security and trustworthiness throughout development and regularly post-deployment, to address evolving threats and ensure ethical standards. The typology of AI threats should build on existing frameworks such as MITRE ATLAS and evolve with the threats and innovation.

Companion resource to the AI RMF - The initiative to establish companion resource to the AI Risk Management Framework (AI RMF) tailored explicitly to AI applications while conducting continuous monitoring and evaluation of AI applications is crucial for proactive risk mitigation and should further incorporate comprehensive risk assessment methodologies, clear risk categorization frameworks building on MITRE ATLAS,, and robust risk mitigation strategies that are supported by automation for compliance and risk posture management - to ensure effectiveness, safety, and adherence to controls. The provision for waiving individual AI applications from certain elements of controls, allows for flexibility in cases where strict adherence may not be feasible or may hinder innovation. However, rigorous justification and alternative risk mitigation measures should accompany this waiver process, similar to the need to provide justification under other established risk frameworks (e.g., ISO/IEC 27001, 27002). Furthermore, the general approach under the RMF should expand the lens of protecting LLM - to protect risks LLM creates as it interacts with the broader threat landscape.

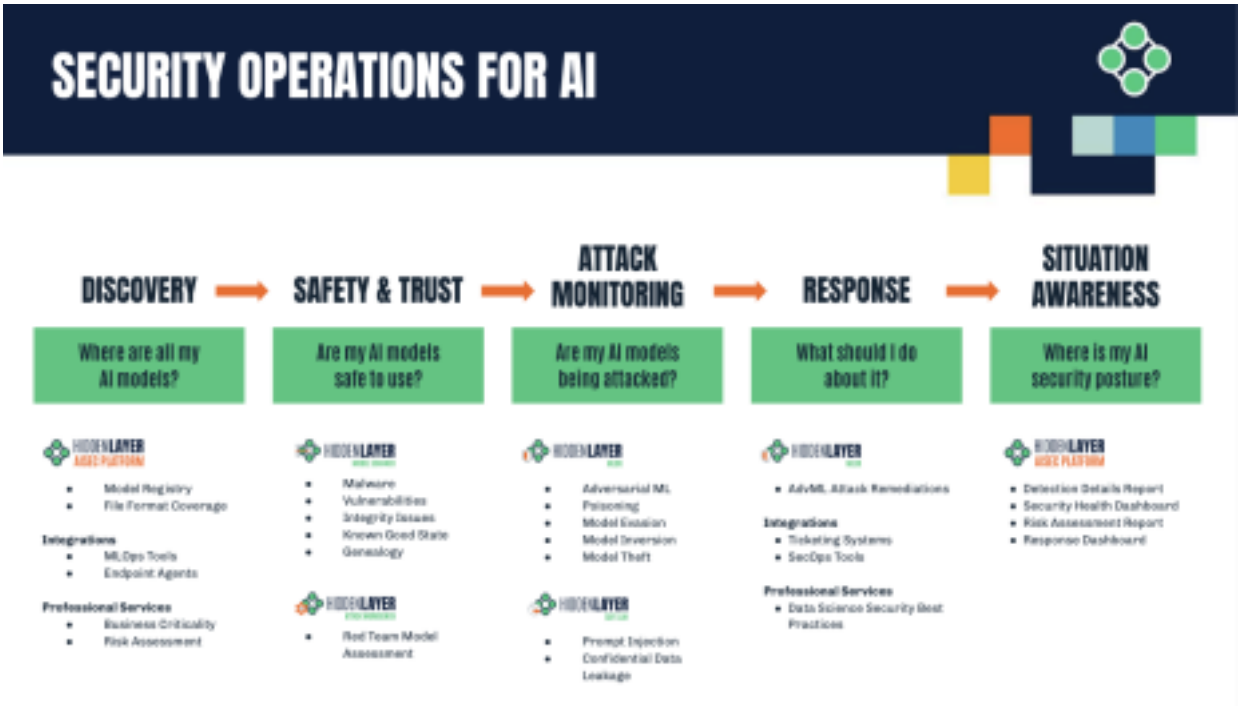
Additionally, fostering open channels for sharing information on AI risks, including non-security issues like bias, underlines the importance of community-wide cooperation in enhancing AI system trustworthiness, guided by frameworks like

NIST's AI Risk Management Framework and CSF 2.0 more broadly, with the focus on coordinated vulnerability disclosure like processes (bug bounty, red-teaming and vulnerability disclosure programs) that span both security and broader AI-harms. That said, recommending such processes should be done in conjunction with allowing, and developing, better protections for third-party testing and via contractual "safe harbors" and legislative protections, for those invited to test these systems.[2]

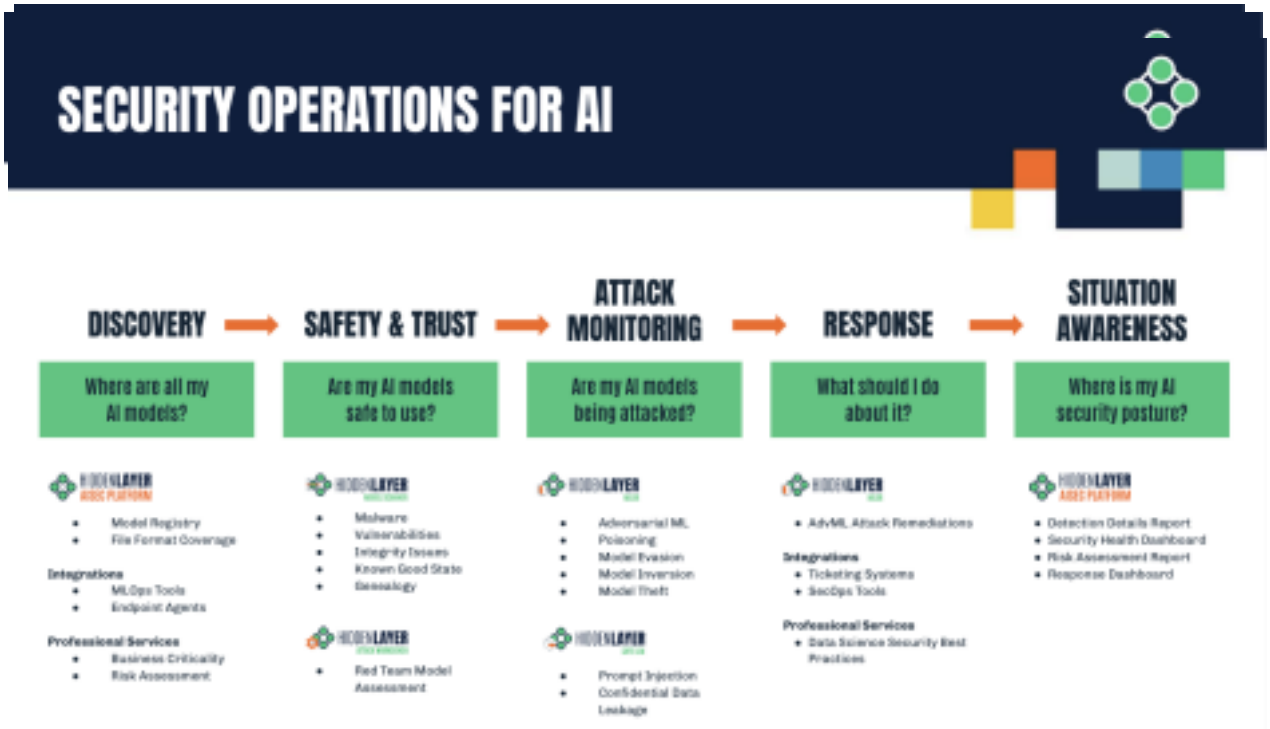
AI system and models development- As mentioned in the context of the OMB AI memo, AI EO and other related publications, agencies must assess potential risks, document stakeholders impacted by AI, and consider failure modes; they should also implement automated tools and technology contextualization techniques to identify and prioritize potential risks associated with AI applications, monitor AI models for drifts in performance, analyze AI models, data sets and maintain data privacy within their AI models and intended use cases to identify potential biases, security vulnerabilities, and ethical concerns.

To further enhance the security of AI technologies, adopting guidelines and best practices that mandate the protection of AI models and their data, including

[2] See, ***OpenPolicy comments on the DMCA 1201 proceeding***, https://downloads.regulations.gov/COLC-2023-0004-0064/attachment_1.pdf. See also HPC comments on this proceeding and DMCA 1201 proceeding.



Openpolicy



implementing solutions to detect adversarial attacks, ensuring data privacy in compliance with regulations such as GDPR and CCPA and emerging AI regulations,, and monitoring model performance is key. Furthermore non-invasive, software-based measures can support transparency, adaptability, with minimal disruption, thereby safeguarding AI models without compromising their functionality or data integrity.

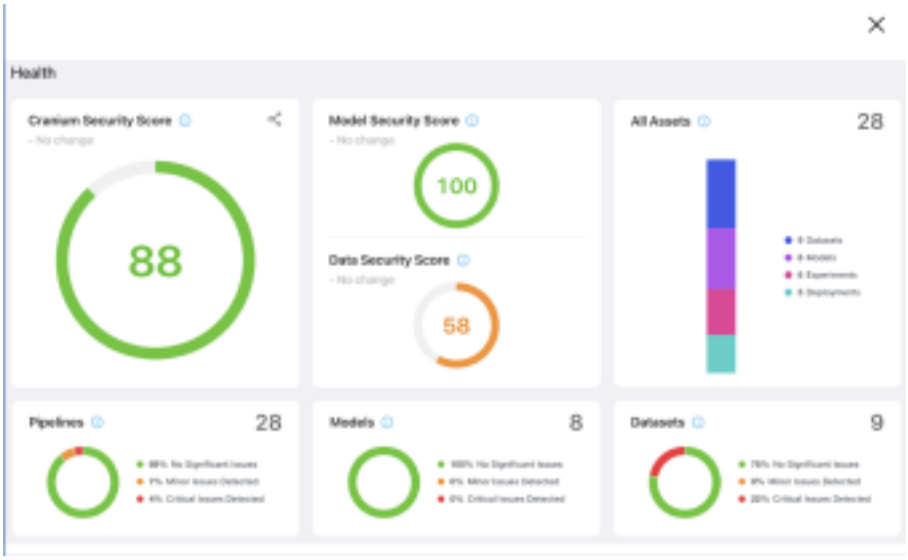
[Potential methodology of enhancing the security of AI across NIST Risk pillars, source: HiddenLayer]

Additionally, developing a comprehensive risk assessment framework should consider both technical and non-technical risks related to AI, including factors such as data privacy, algorithmic bias, fairness, explainability, and potential impacts on underserved communities. Finally, it is critical that model protection controls consider both in training and application approaches and consider the threats poised by using the LLM and different contexts as well as approaches to control which data or unstructured content is fed to the model

Automated risk strategies - A key focus should be given to establishing transparent and automated risk reporting, testing, and mitigation strategies to communicate identified risks and mitigation strategies to relevant stakeholders, including agency leadership, policy experts, and affected communities, to foster trust and accountability in AI cross-sector participation and ecosystem. Measurement of agency posture should be tied to accountability measures to support more robust adoption of controls and foster transparency and interpretability that could be used and elevated through automation of software-based approaches to securing AI models and identifying risks with minimal disruption to existing model workflows or data pipelines.

This approach can include software solutions to support Inventorying of AI assets including datasets, models, experiments, and deployments, consistent with the AI

OMB requirements, mapping of assets into development pipelines and monitoring and alerting for key adversarial threats including, data poisoning, model backdoors, model evasion, model inversion. It further includes providing transparency for compliance reporting and supply chain visibility and sharing.



[AI/model risk measurement in support of compliance and control assessment; Source: Cranium AI]

Putting AI in context – AI and the OT/ IoT landscape – In considering AI standards and controls, it is essential to consider both IT and IoT devices due to their role in digital transformation and inherent security risks as well as botnet enablement. As the global threat landscape increases, IoT devices present vulnerabilities that could compromise data and personal safety, underscoring the urgency for security by design in IoT and OT environments. We encourage the NIST RFI AI team to work closely with their NIST partners focusing on IoT Security, and incorporating controls to address these threats. AI and machine learning advancements and threats, highlight the importance of integrating robust security measures and continuous monitoring of IoT, emphasizing the need for comprehensive standards addressing the security and functionality of interconnected devices in AI ecosystems.

Putting AI in context – AI and the Bot landscape – When considering the full scope of AI challenges and threats, in addition to threats of LLM, there are two components to consider: 1) AI-generated content, and 2) AI autonomous agents that drive live automation and interactions with software systems and people. This activity can serve both legitimate and illegitimate purposes. Legitimate use cases include data analysis, research, training machine learning models, or enhancing use experiences on e-commerce or social media platforms. In contrast, illegitimate use cases involve AI autonomous agents interacting with content or media to hijack trending algorithms, access password-protected systems, scrape sensitive data, or scan

software systems for vulnerabilities. ***In the context of considering the risk of***

synthetic content, NIST should address this interaction between AI-generated content and bot-or other automated methods-elevated attacks. To protect against these threats and differentiate between legitimate and illegitimate AI autonomous agents, there is a need for requirements for both disclosure of and controls around the use of AI autonomous agents in driving interactions with software systems and AI-generated content.

Combating digital threats **-** As AI technologies become more integrated into the digital infrastructure, they can potentially be exploited for malicious purposes, such as bot attacks, fraud, and digital abuse, undermining the RFI's broader goal of ensuring a secure and trustworthy digital environment. Establishing standards prioritizing security and ethical considerations helps safeguard against these threats, preserving the reliability and safety of internet ecosystems for users and businesses alike, an essential proactive approach for fostering innovation while protecting against the exploitation of AI systems.

Secure AI lifecycle - Steps should be taken to ensure transparency and performance of procured AI, including obtaining adequate documentation by evaluating performance claims and developing automated interoperability testing frameworks to evaluate the compatibility of AI solutions with existing systems to ensure seamless integration across different vendors and technologies, considering contracting provisions for continuous improvement. As well as establishing automated data protection monitoring mechanisms and data sharing protocols to track and enforce compliance with data protection regulations and ensure the security and privacy of sensitive data throughout the AI lifecycle. All of this requires a more technology-scaled approach to compliance and measurements.

As the NIST RFI Implementation evolve, we look forward to discussing these proposals with NIST and are available for any questions. We remain excited to collaborate with NIST to increase engagement with innovative companies.

Respectively,

Dr. Amit Elazari, CEO & Co-Founder, OpenPolicy