Openpolicy

June 2, 2024
To: NIST
Via NIST-AI-600-1@nist.gov

OpenPolicy comments on
National Institute of Standards and Technology (NIST) AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

Overview

OpenPolicy appreciates the opportunity to provide feedback on the NIST AI 600-1: Artificial Intelligence Risk Management Framework - Generative Artificial Intelligence Profile. 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 as this framework develops and are ready to offer additional input as necessary. We remain eager to support the implementation of the AI Executive Order and its objectives. We are committed to supporting the use and implementation of this essential document designed to guide organizations designing, developing, deploying, or using AI systems to help manage AI risks and promote trustworthy and responsible development and use, actively engaging with NIST and other implementing agencies alongside our participation in the NISA AI Safety Institute Consortium.

Following NIST's active engagement with government, industry, academia, and the public to understand their needs and develop practical, updated solutions and best practices, we believe that an open and collaborative policymaking dialogue is essential. This approach enhances the implementation of NIST's best practices and ongoing efforts to ensure a secure Gen-AI lifecycle by focusing on the core risks posed by Gen-AI products, which is critical to fostering an appropriate, updated, and scalable risk management framework. The participation of innovative companies, particularly startups specializing in cutting-edge AI security, IoT, and safety solutions, significantly contributes to this effort. Indeed, many technologies used to support the requirements of the AI EO, OMB AI memo, and the NIST RFIs and guidelines referenced in this draft are developed by these innovative companies and startups. These technologies evaluate the measurements, considerations, and development of AI product architecture, deployment strategies, and overarching cybersecurity, data privacy, and secure-by-design perspectives that should guide manufacturers, which OpenPolicy actively collaborates with these communities.

OpenPolicy encourages NIST to identify further opportunities to utilize technology and automation tools to ensure that agency adoption practices align with a proactive, secure, and safe approach to AI. Establishing AI security and risk measurement controls is a vital step, but a scalable strategy is necessary to ensure these controls are effectively implemented. A risk-based approach can only be scaled, given the constantly evolving nature of AI threats, by leveraging solutions to dynamically assess the risk of the system or model in relation to the controls used. Further, OpenPolicy applauds NIST for its leadership in implementing the AI EO requirements, especially satisfied with the emphasis on governance and managing the performance and risks of AI, including AI security, AI red teaming, and automation. A comprehensive approach addressing security across the entire AI lifecycle is crucial. This includes real-time threat detection, monitoring data flows for unauthorized disclosures, and enforcing comprehensive protection policies that foster and introduce proactive measurement crucial to adapting and reacting effectively to the ongoing development of Gen-AI. Such as integrating automated tools with human testers for AI red teaming can uncover nuanced issues such as misinformation and deepfakes, fostering a collaborative environment for thorough evaluation and effective risk mitigation

It has been proven through various standards and best practices endavors, ***governance*** based on documentation alone is insufficient. Although the draft's approach for control measurement is highly consistent with the direction of NIST CSF 2.0, there is a need to include governance supported by automated means and measures. Incorporating advanced technology solutions can significantly bolster information integrity and security. These solutions ensure the secure exchange of sensitive information with encryption and secure access controls. They provide detailed audit trails, real-time monitoring, and compliance with regulations like GDPR and CCPA, which are crucial for managing AI-related data privacy risks. Furthermore, advanced threat detection and real-time response mechanisms help identify and mitigate risks associated with unauthorized data access or breaches which directly correlate with this draft proactive approach to governance and compliance.

For example, as sections ***GV-1.2-001 to GV-1.2-008*** focus on integrating policies and procedures to control risks and ensure accountability, to enhance the framework's ability to manage information integrity, data privacy, and intellectual property risks in a proactively, ongoing manner there is a need to address advanced technology solutions for secure file sharing, governance, compliance, and real-time threat detection which enable protection and detection scalability while facing the growing reach of Gen-AI technology. It should include recommendations for a scalable strategy supported by automated means to ensure the effective implementation of AI security and risk measurement controls which will not

only help in risk management specifically but also in addressing acceptable uses and updating risk tolerances.

While addressing the section on ***Intellectual Property Infringements***, there is a need to address the pivotal role of time and efficiency to ensure the safeguarding of AI models without compromising their functionality. Ensuring transparency and proper documentation of training data sources is essential to mitigate risks related to intellectual property infringements while real-time detection of intellectual property infringements and the use of non-invasive, software-based measures can facilitate that endeavor and prompt this draft objectives. Measurements such as robust data encryption should not solely be mentioned under the ***Approaches for mapping AI technology and legal risks of its components*** but rather be integrated as an ensure tool that all data, including AI training data, is encrypted both at rest and in transit to prevent unauthorized access and data breaches. Implementing granular access controls ensures that only authorized personnel can access sensitive AI data and intellectual property and thus, providing comprehensive audit logs to track data access and usage is critical for ensuring compliance and protecting intellectual property rights.

Further, comprehensive data protection monitoring mechanisms and data sharing protocols are vital to ensure compliance with data privacy regulations throughout the AI lifecycle and thus should also adopt advanced privacy measures and continuous monitoring to detect adversarial attacks and safeguard data privacy effectively. We emphasize the need for interoperability testing frameworks to evaluate the compatibility of AI solutions with existing systems, ensuring seamless integration while maintaining high data privacy standards. The correlation and connection of data privacy regulations such as GDPR and CCPA and robust data protection measures hold the power to enable the secure sharing of sensitive data, Real-time monitoring, and alerts for any suspicious activity or potential data breaches with external parties while maintaining strict privacy controls, ensuring compliance, immediate response, and mitigation.

Furthermore, a holistic approach to addressing LLM threats needs to consider the full model and AI cycle. 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, anomaly, and threat detection. 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 GOVERN 3.2 section addresses***. 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.

Specific comments

MEASURE 2.7: AI system security and resilience – as identified in the MAP function – are evaluated and documented.

Revised version: AI system functionality, capabilities, limitations, safety, security, privacy, effectiveness, suitability, equity, trustworthiness, and resilience – as identified in the MAP function – are evaluated and documented.

MS-2.7-001- Apply established security measures to: Assess risks of backdoors, compromised dependencies, data breaches, eavesdropping, man-in-the-middle attacks, reverse engineering other baseline security concerns; Audit supply chains to identify risks arising from, e.g., data poisoning and malware, software and hardware vulnerabilities, third-party personnel and software; Audit GAI systems, pipelines, plugins and other related artifacts for unauthorized access, malware, and other known vulnerabilities.

Revised version: Apply established security measures to: Assess risks of backdoors, compromised dependencies, data breaches, eavesdropping, man-in-the-middle attacks, autonomous agents, reverse engineering, model theft, AI inference, bypass, extraction, and other baseline security concerns; Audit supply chains to identify risks arising from, e.g., data poisoning and malware, software and hardware vulnerabilities, third-party personnel and software; Audit GAI systems, pipelines, plugins and other related artifacts for unauthorized access, malware, and other known vulnerabilities.

Risks- Data Privacy, Information Integrity, Information Security, Value Chain and Component Integration

MS-2.7-002- Assess the completeness of documentation related to data provenance, access controls, and incident response procedures. Verify GAI system content provenance documentation aligns with relevant regulations and standards.

Risks- Information Integrity, Toxicity, Bias, and Homogenization.

MS-2.7-003- Benchmark GAI system security and resilience related to content provenance against industry standards and best practices. Compare GAI system security features and content provenance methods against industry state-of-the-art.

Revised version: Benchmark GAI system functionality, safety, and trustworthiness including security, transparency, and resilience related to content provenance against industry standards and best practices. Compare GAI system security features and content provenance methods against industry state-of-the-art.

Risks- Information Integrity, Information Security

MS-2.7-004 Conduct user surveys to gather user satisfaction with the AI-generated content and user perceptions of content authenticity. Analyze user feedback to identify concerns and/or current literacy levels related to content provenance.

Risks- Human AI Configuration, Information Integrity

MS-2.7-005 Engage with security experts, developers, and researchers through information-sharing mechanisms to stay updated with the latest advancements in AI security related to content provenance. Contribute findings related to AI system security and content provenance via information-sharing mechanisms, workshops, or publications.

Revised version: Engage with security experts, developers, and researchers through information-sharing mechanisms to stay updated with the latest advancements in AI security related to content provenance, including AI-generated content, data analysis, research, training machine learning models, or enhancing use experiences on e-commerce or social media platforms. Contribute findings related to AI system security and content provenance via information-sharing mechanisms, workshops, or publications.

Risks-Information Integrity, Information Security

MS-2.7-006 Establish measures and evaluate GAI resiliency as part of pre-deployment testing to ensure GAI will function under adverse conditions and restore full functionality in a trustworthy manner.

MS-2.7-007 Identify metrics that reflect the effectiveness of security measures, such as data provenance, the number of unauthorized access attempts, penetrations, or provenance verification.

Revised version: Identify metrics that reflect the effectiveness of security measures, such as data provenance, the number of unauthorized access attempts, inference, bypass, extraction, penetrations, or provenance verification

Risks- Information Integrity, Information Security

MS-2.7-008 Maintain awareness of emergent GAI security risks and associated countermeasures through community resources, official guidance, or research literature.

Risk- Information Security, Unknowns

MS-2.7-009 Measure reliability of content provenance verification methods, such as watermarking, cryptographic signatures, hashing, blockchain, or other content provenance techniques. Evaluate the rate of false positives and false negatives in content provenance, as well as true positives and true negatives for verification.

Revised version: Measure the reliability of content provenance verification methods, such as watermarking, cryptographic signatures, hashing, blockchain, access controls, compliance auditing, model integrity verification or other content provenance techniques. Evaluate the rate of false positives and false negatives in content provenance, as well as true positives and true negatives for verification.

Risk- Information Integrity

MS-2.7-010 Measure the average response time to security incidents related to content provenance, and the proportion of incidents resolved with and without significant impact.

Revised version: Measure the average response real-time to security incidents related to content provenance and analyze the proportion of incidents resolved with and without significant impact.

Risk- Information Integrity, Information Security

MS-2.7-011 Measure the rate at which recommendations from security audits and incidents are implemented related to content provenance. Assess how quickly the AI system can adapt and improve based on lessons learned from security incidents and feedback related to content provenance.

Revised version: Measure the rate at which recommendations from security audits and incidents are implemented related to content provenance. Assess how quickly the AI system can adapt and improve based on lessons learned from security incidents and feedback related to content provenance including confidentiality of sensitive information, sensitive content communications, and AI, Gen AI, and ML models content.

Risk- Information Integrity, Information Security

MS-2.7-012 Monitor and review the completeness and validity of security documentation and verify it aligns with the current state of the GAI system and its content provenance.

Risk- Information Integrity, Information Security, Toxicity, Bias, and Homogenization

MS-2.7-013 Monitor GAI system downtime and measure its impact on operations.

MS-2.7-014 Monitor GAI systems in deployment for anomalous use and security risks.

Risk- Information Security

MS-2.7-015 Monitor the number of security-related incident reports from users, indicating their awareness and willingness to report issues.

Revised version: Monitor the number of security-related incident reports from users and industry databases, indicating their awareness and willingness to report issues.

Risk- Human AI Configuration, Information Security

MS-2.7-016 Perform AI red-teaming to assess resilience against: Abuse to facilitate attacks on other systems (e.g., malicious code generation, enhanced phishing content), GAI attacks (e.g., prompt injection), ML attacks (e.g., adversarial examples/prompts, data poisoning, membership inference, model extraction, sponge examples).

Revised version: Perform AI red-teaming, both red teamers for broad flaws detection and specialists for nuanced issues to assess resilience against: Abuse to facilitate attacks on other systems (e.g., malicious code generation, enhanced phishing content), GAI attacks (e.g., prompt injection), ML attacks (e.g., adversarial examples/prompts, data poisoning, membership inference, model extraction, misinformation and deepfakes sponge examples).

Risk- Information Security, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations

MS-2.7-017 Review deployment approval processes and verify that processes address relevant GAI security risks.

Risk- Information Security

MS-2.7-018 Review incident response procedures and verify adequate functionality to identify, contain, eliminate, and recover from complex GAI system incidents that implicate impacts across the trustworthy characteristics.

MS-2.7-019 Track and document access and updates to GAI system training data; verify appropriate security measures for training data at GAI vendors and service providers.

Revised version: Track, monitor, document access, and updates to GAI system training data; verify appropriate security measures for training data at GAI vendors and service providers.

Risk- Information Security, Value Chain and Component Integration

MS-2.7-020 Track GAI system performance metrics such as response time and throughput under different loads and usage patterns related to content provenance.

Risk- Information Integrity

MS-2.7-021 Track the number of users who have completed security training programs regarding the security of content provenance.

Risk- Human AI Configuration, Information Integrity, Information Security

MS-2.7-022 Verify fine-tuning does not compromise safety and security controls.

Risk- Information Integrity, Information Security, Dangerous or Violent Recommendations

MS-2.7-023 Verify organizational policies, procedures, and processes for treatment of GAI security and resiliency risks.

Risk- Information Security

MS-2.7-024 Verify vendor documentation for data and software security controls.

Risk- Information Security, Value Chain and Component Integration

MS-2.7-025 Work with domain experts to capture stakeholder confidence in GAI system security and perceived effectiveness related to content provenance.

Risk- Information Integrity, Information Security