ode: 3510-13

DEPARTMENT OF COMMERCE

National Institute of Standards and Technology

XRIN: 0693-XA002

Request for Information Regarding Security Considerations for Artificial Intelligence Agents

AGENCY: Center for AI Standards and Innovation (CAISI), National Institute of Standards and Technology (NIST), U.S. Department of Commerce

ACTION: Notice; Request for Information (RFI)

SUMMARY: The Center for AI Standards and Innovation (CAISI), housed within the National Institute of Standards and Technology (NIST) at the Department of Commerce, is seeking information and insights from stakeholders on practices and methodologies for measuring and improving the secure development and deployment of artificial intelligence (AI) agent systems. AI agent systems are capable of taking autonomous actions that impact real-world systems or environments, and may be susceptible to hijacking, backdoor attacks, and other exploits. If left unchecked, these security risks may impact public safety, undermine consumer confidence, and curb adoption of the latest AI innovations. We encourage respondents to provide concrete examples, best practices, case studies, and actionable recommendations based on their experience developing and deploying AI agent systems and managing and anticipating their attendant risks. Responses may inform CAISI's work evaluating the security risks associated with various AI capabilities, assessing security vulnerabilities of AI systems, developing evaluation and assessment measurements and methods, generating technical guidelines and best practices to measure and improve the security of AI systems, and other activities related to the security of AI agent systems.

DATES: Comments containing information in response to this notice must be received on or before [INSERT DATE 60 DAYS FROM THE DATE OF PUBLICATION IN THE FEDERAL

REGISTER], at 11:59 PM Eastern Time. Submissions received after that date may not be considered.

ADDRESSES: Comments must be submitted electronically via the Federal e-Rulemaking Portal.

1. Go to *www.regulations.gov* and enter NIST-2025-0035 in the search field;
2. Click the "Comment Now!" icon, complete the required fields, including the relevant document number and title in the subject field; and
3. Enter or attach your comments.

Additional information on the use of *regulations.gov*, including instructions for accessing agency documents, submitting comments, and viewing the docket is available at: *www.regulations.gov/faq*. If you require an accommodation or cannot otherwise submit your comments via *regulations.gov*, please contact NIST using the information in the FOR FURTHER INFORMATION CONTACT section below.

NIST will not accept comments for this notice by postal mail, fax, or email. To ensure that NIST does not receive duplicate copies, please submit your comments only once. Comments containing references, studies, research, and other empirical data that are not widely published should include copies of the referenced materials.

All relevant comments received by the deadline will be posted at: *https://www.regulations.gov* under docket number NIST-2025-0035 without change or redaction, so commenters should not include information they do not wish to be posted publicly (*e.g.*, personal or confidential business information).

FOR FURTHER INFORMATION CONTACT: For questions about this Request for Information (RFI) contact: Peter Cihon, Senior Advisor, Center for AI Standards and Innovation ((202) 695-5661; peter.cihon@nist.gov). Direct media inquiries to NIST's Office of Public Affairs at (301) 975-2762. Users of telecommunication devices for the deaf, or a text telephone may call the Federal Relay Service toll free at 1-800-877-8339. NIST will make the RFI

available in alternate formats, such as Braille or large print, upon request by persons with disabilities.

SUPPLEMENTARY INFORMATION:

Authority

This RFI advances NIST's activities to support measurement research and development of best practices for artificial intelligence systems, including their safety and robustness to adversarial attacks (15 U.S.C. 278h-1(b)). It is consistent with NIST's functions to, *inter alia*, compile data, provide a clearinghouse of scientific information, and assist industry in improving product quality (15 U.S.C. 272(b-c)).

Background

AI agent systems are capable of planning and taking autonomous actions that impact real-world systems or environments. AI agent systems consist of at least one generative AI model and scaffolding software that equips the model with tools to take a range of discretionary actions. These systems may be more expansive, containing multiple sub-agents with software that orchestrates their interactions. They can be deployed with little to no human oversight. Other terms used to refer to AI agent systems include AI agents and agentic AI. Challenges to the security of AI agent systems may undermine their reliability and lessen their utility, stymieing widespread adoption that would otherwise advance U.S. economic competitiveness. Further, security vulnerabilities may pose future risks to critical infrastructure or catastrophic harms to public safety (*i.e.*, through chemical, biological, radiological, nuclear, and explosive (CBRNE) weapons development and use or other analogous threats).

Deployed AI agent systems may face a range of security threats and risks. Some of these risks are shared with other kinds of software systems, such as exploitable vulnerabilities in authentication mechanisms or memory management processes. This Request for Information, however, focuses instead on the novel risks that arise from the use of machine learning models embedded within AI agent systems. Within this category are: (1) security risks that arise from

adversarial attacks at either training or inference time, when models may interact with potentially adversarial data (*e.g.*, indirect prompt injection) or may be compromised by data poisoning; (2) security risks posed by models with intentionally placed backdoors; and (3) the risk that the behavior of uncompromised models may nonetheless pose a threat to confidentiality, availability, or integrity (*e.g.*, models that exhibit specification gaming or otherwise pursue misaligned objectives). Organizations have begun to implement technical controls, processes, and other mitigations for the security risks posed by their AI agent systems. In some cases, mitigations draw on cybersecurity best practices, including implementing systems according to the principle of least privilege and designing systems with a zero trust architecture. In other cases, risks are addressed with novel approaches, including instruction hierarchy and agent design patterns with trusted models.

NIST conducts research and develops guidelines to promote safe and secure AI innovation and adoption. Research by CAISI technical staff[1] has demonstrated risks of agent hijacking. NIST has also produced resources on this topic including NIST AI 100-2e2025[2] that provides a taxonomy of attacks and mitigations in adversarial machine learning generally; the NIST AI Risk Management Framework,[3] which describes and discusses "secure and resilient" AI and includes subcategories for security assessment within the Measure function; NIST's companion Risk Management Framework: Generative AI Profile,[4] which provides further context and considerations for "information security" and associated risks with generative AI, applicable to this RFI; and NIST AI 800-1[5] that provides guidelines for AI developers to manage risks including the misuse of AI agent systems for offensive cybersecurity operations. In addition, NIST SP 800-218A[6] provides a profile for the secure development of generative AI, and NIST SP 800-53[7] provides a glossary of relevant terms and a catalog of security and privacy controls for information systems generally.

Request for Information

This RFI seeks information that can support secure innovation and adoption of AI agent

systems. It invites stakeholders—particularly AI agent developers, deployers, and computer security researchers—to share insights on the secure development and deployment of AI agent systems. Such information should be scoped to the security of AI agent systems capable of taking actions that affect external state, *i.e.*, persistent changes outside of the AI agent system itself. Unless contextualized to impact the security of agent systems directly, this RFI does not seek general information on generative AI security, insights on practices for AI chatbots or retrieval-augmented generation systems that are not orchestrated to act autonomously, or feedback on the misuse of AI agent systems to carry out cyberattacks.

NIST is requesting that respondents provide information on the topics below. NIST has provided this non-exhaustive list of topics and accompanying questions to guide respondents, and the submission of any relevant information germane to the subject but that is not included in the list of topics below is also encouraged. NIST will consider all relevant comments received during the public comment period. Respondents need not address all questions in this RFI, though all responses should specify which questions are being answered. For respondents with limited bandwidth, please prioritize questions 1(a), 1(d), 2(a), 2(e), 3(a), 3(b), 4(a), 4(b), and 4(d). All relevant responses that comply with the requirements listed in the **DATES** and **ADDRESSES** sections of this RFI will be considered.

1. Security Threats, Risks, and Vulnerabilities Affecting AI Agent Systems

(a) What are the unique security threats, risks, or vulnerabilities currently affecting AI agent systems, distinct from those affecting traditional software systems?

(b) How do security threats, risks, or vulnerabilities vary by model capability, agent scaffold software, tool use, deployment method (including internal vs. external deployment), hosting context (including components on premises, in the cloud, or at the edge), use case, and otherwise?

(c) To what extent are security threats, risks, or vulnerabilities affecting AI agent systems creating barriers to wider adoption or use of AI agent systems?

(d) How have these threats, risks, or vulnerabilities changed over time? How are they likely to evolve in the future?

(e) What unique security threats, risks, or vulnerabilities currently affect multi-agent systems, distinct from those affecting singular AI agent systems?

2. Security Practices for AI Agent Systems

(a) What technical controls, processes, and other practices could ensure or improve the security of AI agent systems in development and deployment? What is the maturity of these methods in research and in practice? Categories may include:

i. Model-level controls, such as measures to enhance model robustness to prompt injections;

ii. Agent system-level controls, such as prompt engineering, data or tool restrictions, and continuous monitoring methods;

iii. Human oversight controls, such as approvals for consequential actions, management of sensitive and untrusted data, network access permissions, or other controls.

(b) To what degree, if any, could the effectiveness of technical controls, processes, and other practices vary with changes to model capability, agent scaffold software, tool use, deployment method (including internal vs. external deployment), use case, use in multi-agent systems, and otherwise?

(c) How might technical controls, processes, and other practices need to change, in response to the likely future evolution of AI agent system capabilities or of the threats, risks, or vulnerabilities facing them?

(d) What are the methods, risks, and other considerations relevant for patching or updating AI agent systems throughout the lifecycle, as distinct from those affecting both traditional software systems and non-agentic AI?

(e) Which cybersecurity guidelines, frameworks, and best practices are most relevant to the security of AI agent systems?

i. What is the extent of adoption by AI agent system developers and deployers of these relevant guidelines, frameworks, and best practices?

ii. What are impediments, challenges, or misconceptions about adopting these kinds of guidelines, frameworks, or best practices?

iii. Are there ways in which existing cybersecurity best practices may not be appropriate for the security of AI agent systems?

3. Assessing the Security of AI Agent Systems

(a) What methods could be used during AI agent systems development to anticipate, identify, and assess security threats, risks, or vulnerabilities?

i. What methods could be used to detect security incidents after an AI agent system has been deployed?

ii. How do these align (or differ) from traditional information security practices, including supply chain security?

iii. What is the maturity of these methods in research and applied use?

iv. What resources or information would be useful for anticipating, identifying, and assessing security threats, risks, or vulnerabilities?

(b) Not all security threats, risks, or vulnerabilities are necessarily applicable to every AI agent system; how could the security of a particular AI agent system be assessed and what types of information could help with that assessment?

(c) What documentation or data from upstream developers of AI models and their associated components might aid downstream providers of AI agent systems in assessing, anticipating, and managing security threats, risks, or vulnerabilities in deployed AI agent systems?

i. Does this data or documentation vary between open-source and closed-source AI models and AI agent systems, and if so, how?

ii. What kinds of disclosures (if made mandatory or public) could potentially create new vulnerabilities?

iii. How should such, if any, disclosures be kept secure between parties to protect system integrity?

(d) What is the state of practice for user-facing documentation of AI agent systems that support secure deployment?

4. Limiting, Modifying, and Monitoring Deployment Environments

(a) AI agent systems may be deployed in a variety of environments, *i.e.*, locations where the system's actions take place. In what manner and by what technical means could the access to or extent of an AI agent system's deployment environment be constrained?

(b) How could virtual or physical environments be modified to mitigate security threats, risks, or vulnerabilities affecting AI agent systems? What is the state of applied use in implementing undoes, rollbacks, or negations for unwanted actions or trajectories (sequences of actions) of a deployed AI agent system?

(c) What is the state of managing risks associated with interactions between AI agent systems and counterparties? Practices, their adoption, and their relative maturity may differ according to the counterparty in the interaction, including:

i. Interactions with humans who are not using the AI agent system directly;

ii. Interactions with digital resources, including web services, servers, and legacy systems;

iii. Interactions with mechanical systems, machinery, or Internet-of-Things (IoT);

iv. Interactions with authentication mechanisms, operating system access, source code access, or similar network-level access vectors;

v. Interactions with other AI agent systems.

(d) What methods could be used to monitor deployment environments for security threats, risks, or vulnerabilities?

i. What challenges exist to deploying traditional methods of monitoring threats, risks, or vulnerabilities?

ii. Are there legal and/or privacy challenges to monitoring deployment environments for security threats, risks, or vulnerabilities?

iii. What is the maturity of these methods in research and practice?

(e) Are current AI agent systems widely deployed on the open internet, or in otherwise unbounded environments? How could the volume of traffic be tracked on the open internet or in otherwise unbounded environments over time?

5. Additional Considerations

(a) What methods, guidelines, resources, information, or tools would aid the AI ecosystem in the rapid adoption of security practices affecting AI agent systems and promoting the ecosystem of AI agent system security innovation?

(b) In which policy or practice areas is government collaboration with the AI ecosystem most urgent or most likely to lead to improvements in the state of security of AI agent systems today and into the future?

(c) In which critical areas should research be focused to improve the current state of security practices affecting AI agent systems?

i. Where should future research be directed in order to unlock the benefits of adoption of secure and resilient AI agent systems?

ii. Which research approaches should be prioritized to advance the scientific understanding and mitigation of security threats, risks, and vulnerabilities affecting AI agent systems?

(d) How are other countries addressing these challenges and what are the benefits and drawbacks of their approaches?

(e) Are there practices, norms, or empirical insights from fields outside of artificial intelligence and cybersecurity that might benefit our understanding or assessments of the security of AI agent systems?

Footnotes

1. Technical Blog: Strengthening AI Agent Hijacking Evaluations, *https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations.*

2. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST AI 100-2e2025), *https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf.*

3. Artificial Intelligence Risk Management Framework (NIST AI 100-1), *https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.*

4. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), *https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.*

5. Managing Misuse Risk for Dual-Use Foundation Models (NIST AI 800-1 2pd), *https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf.*

6. Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile (NIST SP 800-218), *https://csrc.nist.gov/pubs/sp/800/218/a/final*.

7. Security and Privacy Controls for Information Systems and Organizations (NIST SP 800-53 Rev. 5), *https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final*.

Alicia Chambers,
NIST Executive Secretariat.
[FR Doc. 2026-00206 Filed: 1/7/2026 8:45 am; Publication Date: 1/8/2026]