PUBLIC COMMENT: FOUNDATION FOR DEFENSE OF DEMOCRACIES

U.S. Department of Commerce
National Institute of Standards and Technology, Center for AI Standards and Innovation

Regarding Security Considerations for Artificial Intelligence Agents

XRIN 0693-XA002

AUTHORS

Leah Siskind
Director of Impact and AI Research Fellow

Dr. Georgianna Shea
Chief Technologist, FDD's Center on Cyber and Technology Innovation and its Transformative Cyber Innovation Lab

Marina Chernin
Research Intern

Washington, DC
March 9, 2026

FDD
FOUNDATION FOR DEFENSE OF DEMOCRACIES

www.fdd.org

Introduction

The emergence of agentic artificial intelligence (AI) represents an inflection point for the federal government. If integrated properly, the technology has the capacity to expand productivity, boost inter-agency coordination, and dramatically accelerate bureaucratic throughput. However, if agentic AI is deployed without due consideration for its potential dangers, it risks undermining public trust and eroding confidence in institutions it might have served.

AI agent systems are software programs that act autonomously to achieve specific goals. They typically consist of one or more large language models (LLMs) embedded within scaffolding software that equips them with tools to plan and execute multi-step tasks — including taking actions on external systems — without continuous human oversight. It is this capacity for autonomous action, rather than the underlying model itself, that distinguishes agent systems from conventional AI applications. That capacity for autonomous action makes agentic AI powerful and uniquely challenging to secure. Because the attack surface includes both systems as well as the agents themselves, existing frameworks do not fully address the national security vulnerabilities AI agents introduce. As federal adoption of agentic AI grows, adversaries can repurpose attacks that have already proven effective against conventional AI, but with consequences that extend beyond manipulated outputs to manipulated actions taken on behalf of the federal government.

At the same time, the efficiency and scalability of agentic AI offer significant operational benefits for federal agencies. The policy challenge is not whether to deploy these systems, but how to do so in a manner that manages risk systematically. The United States has navigated analogous transitions before. The integration of networked computing into government operations generated hard-won lessons about access control, supply chain integrity, and attribution that are directly applicable here. The most transferable of these is the prioritization of transparency, auditability, and accountability as foundational design requirements rather than compliance afterthoughts.

Strategic Risks

There are attack vectors related to the adoption of AI agents, including backdoor attacks, indirect prompt injection, and data poisoning. Russia, China, and Iran have already embraced some of these techniques in the context of regular LLMs. The ability of agentic AI to amplify both impact and stealth, however, will increase the potency and scale of adversarial attacks if not properly accounted for and protected against.

The primary concerns about federal agentic AI adoption revolve around vulnerabilities to adversarial manipulation, cascading failures within interconnected systems, and lack of accountability.

The first potential vulnerability, backdoor attacks, involves manipulating models to behave normally under standard conditions while producing incorrect outputs or taking unauthorized actions when specifically triggered. A trigger could be a phrase, pattern, or even an invisible code embedded in external inputs. For example, federal agencies process enormous amounts of external communications like public submissions and contractor documents. Any of those could carry a trigger.

Chinese state-sponsored hacker groups have employed this technique extensively, designing their operations to be hard to trace back to Beijing. BRICKSTORM is a recent example; state actors from the People's Republic of China created a backdoor for widely used enterprise infrastructure and it remained undetected for a long period of time.[1]

Agentic AI introduces new concerns. First, adversaries can now embed backdoors in the agents themselves rather than in peripheral systems. In February, security firm Koi.ai analyzed ClawHub, the open-source marketplace for AI agent skills and found 824 unauthorized or harmful capabilities (known as malicious skills), out of approximately 10,700.[2] These seemingly innocuous bits of code secretly opened backdoors for their creators which granted them access to everything the bot could interact with. Any users who downloaded them had their systems compromised, exposed to both espionage and external control.

The second and potentially more important concern is called prompt injection, where LLMs are fed malicious instructions to force them to follow their attackers' commands and ignore original programming guidelines. Analysts have found that some currently deployed agents have been conducting prompt injections against others, compelling them to delete their own accounts, run financial manipulation schemes, establish false authority, or spread jailbreak content.[3] The UK's National Cyber Security Center concluded that, because there is no clear delineation between instructions and data for an LLM, it may be impossible to fully excise prompt injections once introduced into a system.[4] Unlike traditional cyberattacks, prompt injection requires no access to government systems. A malicious instruction embedded in an external document or an email is sufficient. By the time the compromise is detected, the damage would already be underway.

In the context of agents, this risk is heightened, as OpenAI itself admitted in December 2025.[5] With access to personal systems, such as email or bank accounts, agents with bad instructions can do far more significant damage. While this method is new, its principles are an AI-adaptation of a much older Russian concept: reflexive control.[6] This is a Cold War-era tactic of psychological subversion, in which one tricks their target into acting against their own interests while believing they are acting in favor of them.

---

[1] Cybersecurity and Infrastructure Security Agency, National Security Agency, and Canadian Centre for Cyber Security, "BRICKSTORM Backdoor," December 4, 2025. (https://www.cisa.gov/news-events/analysis-reports/ar25-338a)

[2] Alex Yomtov and Oren Yomtov, "ClawHavoc: 341 Malicious Clawed Skills Found by the Bot They Were Targeting," *Koi*, February 1, 2026. (https://www.koi.ai/blog/clawhavoc-341-malicious-clawedbot-skills-found-by-the-bot-they-were-targeting)

[3] Kristian McCann, "State-Sponsored Hackers Turn Google's AI Into Full-Spectrum Attack Assistant," *UC Today*, February 12, 2026. (https://www.uctoday.com/security-compliance-risk/state-sponsored-hackers-turn-googles-ai-into-full-spectrum-attack-assistant)

[4] Alex Scroxton, "NCSC warns of confusion over true nature of AI prompt injection," *Computer Weekly*, December 8, 2025. (https://www.computerweekly.com/news/366636155/NCSC-warns-of-confusion-over-true-nature-of-AI-prompt-injection)

[5] Greg Otto, "OpenAI says prompt injection may never be 'solved' for browser agents like Atlas," *CyberScoop*, December 30, 2025. (https://cyberscoop.com/openai-chatgpt-atlas-prompt-injection-browser-agent-security-update-head-of-preparedness)

[6] Maria Snegovaya, "Putin's Information Warfare in Ukraine: Soviet Origins of Russia's Hybrid Warfare," *Institute for the Study of War*, September 21, 2015. (https://understandingwar.org/research/russia-ukraine/putins-information-warfare-in-ukraine-soviet-origins-of-russias-hybrid-warfare-2)

Along the same lines, data poisoning is a method of incapacitating AI systems by providing them with inaccurate or harmful data that shapes their processing and behaviors. Again, Russia has deployed reflexive control methods to shape information environments, change LLM perceptions of issues related to foreign affairs, and thus influence the thinking (and voting) of their users by deploying mass disinformation networks such as "Pravda."[7] In the context of AI agents that autonomously search, retrieve, summarize, and act on information, data poisoning creates a compounding risk: poisoned sources can be amplified, cited, and reintroduced into new knowledge systems, reinforcing the original manipulation.

Where Russia seeks to manipulate what AI systems believe, China seeks to control what they can do. China's "system of systems" warfare doctrine can be construed as their strategy for battlefield AI, but it is simply a doctrine that applies to any high-value decision-making infrastructure, including federal agentic systems.[8] A federal government deploying agentic AI across key operations *is* command-and-control infrastructure. Attacks such as BRICKSTORM, supply chain infiltration, and malicious skills are not random opportunistic attacks — they are consistent with a doctrine that prioritizes degrading an adversary's decision-making systems before a conflict begins.

None of the threats discussed above are unique to agentic AI systems. What changes in a federal context is the scale of the challenge. The difficulty of pinpointing precisely what agents are doing once deployed, who or what they interact with, or where each skill, line of code, or software package comes from, is immense. According to cybersecurity experts, a unique added challenge of agentic AI is multi-agent interaction risk.[9] Unlike traditional software systems, agents are increasingly being deployed in contexts where they interact not with humans, but with other autonomous systems. They can even "spin up" new agents through a process called "self-replication."[10] When such interactions occur, it substantially complicates attribution to, and oversight of, individual agents and, by extension, their owners.

In multi-agent environments, a single compromise does not stay contained — corrupted instructions, poisoned data, and hijacked workflows propagate across interconnected systems faster than any human oversight mechanism can detect or interrupt. With little ability to locate the source of, or attribute blame for, any individual bad action, federal agencies, if they choose to deploy AI agents across key operations, risk not only dangerous dysfunction, but also highly visible operational failures that erode institutional credibility and public confidence.

Learning Lessons From Successful Deployment of Cyber

---

[7] Valentin Chatelet, "Exposing Pravda: How pro-Kremlin forces are poisoning AI models and rewriting Wikipedia," *Atlantic Council*, April 18, 2025. (https://www.atlanticcouncil.org/blogs/new-atlanticist/exposing-pravda-how-pro-kremlin-forces-are-poisoning-ai-models-and-rewriting-wikipedia)

[8] Jeffrey Engstrom, "Systems Confrontation and System Destruction Warfare: How the Chinese People's Liberation Army Seeks to Wage Modern Warfare," *The RAND Corporation*, February 1, 2018. (https://www.rand.org/pubs/research_reports/RR1708.html)

[9] Benjamin Klein, Charlie Lewis, Rich Isenberg, Dante Gabrielli, Helen Mollering, Raphael Engler, and Vincent Yuan, "Deploying agentic AI with safety and security: A playbook for technology leaders," *McKinsey & Company*, October 16, 2025. (https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/deploying-agentic-ai-with-safety-and-security-a-playbook-for-technology-leaders)

[10] "Self-Replicating AI Agents — The Rise of AI That Builds AI," *The Agentics Co*. July 28, 2025. (https://theagentics.co/insights/self-replicating-ai-agents-the-rise-of-ai-that-builds-ai)

The United States is not without recourse. It has one of the most mature cybersecurity ecosystems in the world, built over decades through National Institute of Standards and Technology policy, agency-specific security requirements, sector guidance, and operational practice — and agentic AI, as software, can be addressed here in the same way. The U.S. ecosystem must be the starting point for thinking through how to regulate agentic AI. The task is not to invent a new security philosophy, but to identify where existing frameworks fall short and update them for a world where software acts autonomously.

Existing federal cybersecurity frameworks were designed for deterministic software — systems that execute predefined instructions and nothing more. Agentic AI, which makes decisions, invokes tools, and acts autonomously, does not fit those assumptions. Three foundational NIST frameworks require targeted adaptations. NIST SP 800-160 was built for deterministic software that does exactly what it was programmed to do. It has no framework for software that decides independently what actions to take, no mechanism for controlling what an agent is allowed to do when it reaches for an external tool, and no way to verify safe behavior in systems whose outputs are inherently variable.[11]

NIST SP 800-218, the Secure Software Development Framework (SSDF), was extended for generative AI model development through NIST SP 800-218A, but that profile explicitly stops at the model itself — it does not cover how agentic systems are deployed or operated, leaving serious gaps in adversarial robustness testing, secure design for the tool layer, and supply chain controls for model weights and training data.[12]

NIST SP 800-53 contains parallel control gaps across the Access Control (AC), Identification and Authentication (IA), Audit and Accountability (AU), and Supply Chain Risk (SR) families that leave agentic systems without adequate runtime integrity, identity, provenance, or supply chain protections.[13] NIST SP 800-53 assumes a user can log and attribute actions to specific actors, but in a multi-agent ecosystem where agents are replicating and creating new agents, attribution can be very difficult. By identifying and providing guidance for these currently unaddressed questions, the government can ensure safe adoption of agentic AI that builds public accountability and trust.

Even as the cybersecurity community addresses the gaps in these standards, it also needs a shift in framing: rather than cataloguing vulnerabilities and patching them one by one, security design should start with considering the outcomes that cannot be permitted under any circumstances. Most agentic failures are not model failures. They are authority failures, in which the agent was permitted to do something it should never have been allowed to do. Because agents operate autonomously and at machine speed, failures can propagate across every connected system before any human can detect them. The core security question is therefore: under any failure

[11] National Institute of Standards and Technology, "Engineering Trustworthy Secure Systems," November 2022. (https://csrc.nist.gov/pubs/sp/800/160/v1/r1/final)
[12] National Institute of Standards and Technology, "Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities," February 2022. (https://csrc.nist.gov/pubs/sp/800/218/final); National Institute of Standards and Technology, "Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile," July 2024. (https://csrc.nist.gov/pubs/sp/800/218/a/final)
[13] National Institute of Standards and Technology, "Security and Privacy Controls for Information Systems and Organizations," September 2020. (https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final)

scenario, what is the worst this agent could do, and have we engineered that outcome out of the system?

Idaho National Laboratory's (INL) Consequence-Driven Cyber-Informed Engineering (CCE) methodology offers a way to answer that question by beginning not with vulnerabilities, but with unacceptable outcomes.[14] CCE's logic applies to the full spectrum of agentic risk, whether it be deliberate attack, accidental misuse, misconfiguration, poorly scoped permissions, or goal overreach.

From any starting point, the engineering response is the same: identify the highest-consequence outcomes, map the pathways that enable them, and engineer mitigations that eliminate those outcomes by design. CCE significantly reduces the extent of catastrophic agentic AI failures and is a powerful tool for clearly defined scenarios, but has its limitations. It should be paired with robust monitoring and human oversight to thwart the risks that can't be defined in advance.

Recommendations

1. **Update NIST Systems Engineering and Development Standards.** NIST should update NIST SP 800-160 and NIST SP 800-218 to account for agentic AI across the full system life cycle, from how agentic systems are built to how they are authorized and operated. These updates should establish minimum engineering requirements for action authority, tool invocation security, agent change control, and operational containment, which would give federal agencies a consistent engineering baseline for deploying agentic systems without introducing unacceptable operational risk. Federal procurement requirements should mandate compliance with these updated standards and, similar to existing least-privilege requirements, incorporate as a formal procurement standard the Open Worldwide Application Security Project (OWASP) Agentic Top 10 principle of least agency, which says that agents should receive only the minimum autonomy required for their authorized task.[15]

2. **Establish Agentic AI Testing Standards and Infrastructure.** NIST SP 800-53A, the assessment procedures companion to NIST SP 800-53, must be updated to include agentic-specific evaluation procedures, and federal penetration testing scopes must be explicitly extended to cover the agentic attack surface, including prompt injection, tool abuse, memory poisoning, and multi-agent lateral movement.[16] Testing should be required at initial authorization and repeated following significant model updates and configuration changes. The Center for AI Standards and Innovation should establish a national agentic AI testing capability, with the Department of Energy national laboratories as natural partners, providing shared adversarial evaluation resources for agencies that lack in-house capacity and generating the empirical case study base that

[14] "Consequence-Driven Cyber-Informed Engineering," *Idaho National Laboratory*, accessed February 26, 2026. (https://inl.gov/national-security/cce)
[15] OWASP Gen AI Security Project, "OWASP Top 10 for Agentic Applications for 2026," *Open Worldwide Application Security Project*, December 9, 2025. (https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026)
[16] National Institute of Standards and Technology, "Assessing Security and Privacy Controls in Information Systems and Organizations," January 2022. (https://csrc.nist.gov/pubs/sp/800/53/a/r5/final)

continuously improves MITRE Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS) and informs standards development.

3. **Accelerate the Control Overlays for Securing AI Systems (COSAiS) Initiative.** NIST's COSAiS project develops tailored NIST SP 800-53 control overlays — rules for specific AI use cases that are not yet covered by the broader statute.[17] It is the right vehicle for closing agentic control gaps. As of early 2026, the agentic use case overlays remain in development while federal deployments are already underway. NIST has to prioritize the single-agent and multi-agent overlays and publish interim compensating control guidance for agencies that cannot wait for final publication.

4. **Expand MITRE ATLAS.** MITRE ATLAS must be developed beyond its classical machine-learning origins to cover the potential adversarial uses of AI agents, including agentic kill-chain tactics, reasoning-layer attacks, and multi-agent lateral movement. Federal research and development investment through Defense Advanced Research Projects Agency and National Science Foundation should support this work.

5. **Launch the AI Information Sharing and Analysis Center (AI-ISAC).** The Cybersecurity and Infrastructure Security Agency should immediately stand up the AI-ISAC referenced in the White House AI Action Plan.[18] It should be structured around consequence-informed incident reporting that captures impact chains, rather than just attack indicators. Until launch, existing ISACs should add mandatory agentic AI incident-reporting categories to begin building the collective intelligence that enables cross-sector defense.

6. **AI Bill of Materials.** For agentic AI to be trustworthy and auditable, transparency into system composition is essential. Software bills of materials (SBOMs) serve as ingredient lists for software products, providing visibility into their components and contributing to safety by design. The same logic must extend to AI. Understanding the provenance of data, model lineage, and third-party integrators will help government tech leaders make informed procurement decisions about models with opaque or foreign-influenced training. For example, an AIBOM would help protect agencies from acquiring software trained on datasets with significant Chinese state influence which would already be shaped by censorship and propaganda. The Commerce Department should require all agencies to require an AI bill of materials as condition of procurement for AI vendors.

7. **Humans in the loop.** Comprehensive agentic AI oversight requires both structure and people. Agentic AI risks will impact all of the federal government, so oversight and coordination cannot be limited to specific agencies. A challenge of this magnitude requires a whole-of-government response. The Office of Science and Technology Policy at the White House should manage an inter-agency task force that will share best practices, alert agencies of developing risks, and harmonize practices across departments.

[17] National Institute of Standards and Technology, "SP 800-53 Control Overlays for Securing AI Systems (COSAiS)," July 10, 2025. (https://csrc.nist.gov/projects/cosais)
[18] The White House, "Winning the Race: America's AI Action Plan," July 2025. (https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf)

To maximize the impact of this task force, all agencies must invest in in-house AI expertise. Staffing them with industry experts and maintaining a robust pipeline of tech talent is imperative for this critical moment. Vital government tech offices like the U.S. Digital Service, 18F, and the presidential innovation fellowship need to be reconstituted with a sharpened focus on agentic AI. Lastly, while hiring AI experts into government is necessary, no internal team can match the collective ingenuity of thousands of independent security researchers actively looking for problems. To stay nimble and dynamic, the government should leverage the strengths of the security researcher or "white hat hacker" community, by conducting continuous community-based testing like "bug bounties."

Conclusion

The security challenges posed by agentic AI are not hypothetical. Federal agencies already have the capacity to deploy these systems, adversaries are already probing them, and the frameworks meant to govern them have not kept pace.[19] The recommendations outlined above are not a call to slow adoption. They are a call to make adoption sustainable by ensuring that the infrastructure of oversight, testing, and accountability is built alongside the technology rather than after the fact.

The United States has done this before. The cybersecurity frameworks that now anchor federal IT security were built incrementally, through exactly the kind of standards updates, testing infrastructure, and cross-agency coordination proposed here. The difference is that agentic AI operates at a speed and scale that compresses the timeline for getting it right. Gaps that might have taken years to exploit in prior technology transitions can be leveraged in this one far more quickly.

That urgency is compounded by the strategic context. Russia, China, and Iran are not passive observers of this transition. They have demonstrated both the intent and the capability to exploit vulnerabilities in AI systems, and agentic architectures give them new and potent vectors to do so. Backdoors embedded in agent skills, prompt injections that persist across network rebuilds, and poisoned data that compounds through autonomous retrieval cycles are not theoretical attack scenarios. They are extensions of techniques already in use, applied to a more consequential target. A federal government that deploys agentic AI without closing existing control gaps is one that has handed adversaries a significant and unnecessary advantage.

The window to establish sound foundations is open, but it will not remain so. Acting now, through targeted and engineering-based updates to existing frameworks and by building the human expertise required to implement and oversee them, is the most practical and secure path forward.

[19] Chris Barry, "Accelerating AI adoption for the US government," *Microsoft*, September 2, 2025. (https://blogs.microsoft.com/blog/2025/09/02/accelerating-ai-adoption-for-the-us-government)