Federal Agentic AI Security: NIST's Emerging Standards Initiative

From RFI to SP 800-53 Overlays: What the Federal Framework-in-Progress Means for Enterprise AI Deployments

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-30

Key Takeaways

- NIST's Center for AI Standards and Innovation (CAISI) formally launched the **AI Agent Standards Initiative** on February 17, 2026, establishing the first US government program dedicated explicitly to interoperability and security standards for agentic AI systems, distinct from AI safety evaluation programs under prior executive orders.
- The National Cybersecurity Center of Excellence (NCCoE) published a concept paper in February 2026 proposing to adapt existing identity and authorization frameworks for AI agents—addressing a critical gap in how organizations authenticate and govern agents that act autonomously on behalf of users.
- NIST's empirical research from January 2025 demonstrated that novel attack strategies against AI agents achieved an **81% success rate** in red-team exercises, compared to 11% against baseline defenses—underscoring the urgency of the standards work now underway [1].
- The most technically specific forthcoming guidance—**COSAiS SP 800-53 control overlays** for both single-agent and multi-agent AI systems—remains in development as of March 2026.
- OMB Memoranda M-25-21 and M-25-22 (April 2025) provide actionable federal governance and procurement requirements that apply to many agentic AI deployments today under the "High-Impact AI" classification, even absent agentic-specific regulations.
- Critical gaps persist: no standalone federal agentic AI security standard exists, no FAR clause specifically governs AI agent procurement, and MITRE ATT&CK for Enterprise/ATLAS does not yet cover agentic attack patterns such as multi-agent lateral movement and reasoning-layer manipulation.

Background

The deployment of agentic AI—AI systems capable of autonomous action, tool use, environmental interaction, and multi-step task completion—has outpaced the regulatory and standards frameworks designed to govern it. Where earlier AI deployments functioned primarily as decision-support tools requiring human review at each step, modern AI agents can browse the web, execute code, query databases, send communications, call APIs, and delegate subtasks to other agents with minimal human

intervention. This architectural shift transforms the security threat model fundamentally: an AI agent is not a system that only produces outputs for human review, but one that takes actions with real-world consequences.

The federal government has recognized this distinction, even as its regulatory response remains incomplete. The Biden-era Executive Order 14110 on AI safety, revoked by EO 14179 in January 2025, had begun to establish safety evaluation requirements for frontier AI systems, but neither addressed agentic deployment architectures specifically [2]. The cited policy documents suggest that the current administration's approach—articulated through EO 14179, the White House AI Action Plan of July 2025, and a December 2025 executive order on national AI policy—prioritizes accelerating AI adoption and removing regulatory barriers, with security framing oriented primarily toward preventing technology theft and foreign adversary exploitation rather than establishing domestic deployment requirements for autonomous AI systems [3][4]. EO 14179 does not establish domestic deployment security requirements for autonomous AI systems.

Within this policy context, NIST has emerged as the primary locus of agentic AI security standards work at the federal level, with CISA independently active in parallel workstreams. Through CAISI and the NCCoE, NIST has launched a coordinated set of initiatives since mid-2025 that collectively represent the most technically substantive effort to define what secure agentic AI deployment looks like. Understanding these initiatives—what they cover, what they leave unaddressed, and what their timelines imply for organizations deploying agents today—is essential context for any enterprise security or compliance program.

Security Analysis

The NIST AI Agent Standards Initiative

On February 17, 2026, NIST's CAISI announced the AI Agent Standards Initiative, framing it as a response to both the rapid proliferation of commercial agentic AI deployments and the emerging geopolitical competition in AI standards-setting [5]. The initiative is organized around three strategic pillars: industry-led standards development with US leadership in international bodies such as ISO/IEC JTC 1; community-led open-source protocol development for agents, co-invested with NSF; and fundamental research in AI agent security, identity infrastructure, and interoperability evaluation methodologies.

The initiative built on a Request for Information (RFI) published in the Federal Register on January 8, 2026 (docket NIST-2025-0035), which solicited ecosystem perspectives on the current threat landscape for AI agent systems, existing mitigations and their gaps, measurement methodologies for

agent security, and best practices for secure development [6]. The comment period closed March 9, 2026, and substantive responses were received from the OpenID Foundation, the Foundation for Defense of Democracies (FDD), and commercial AI providers including Perplexity. The FDD's March 9 submission explicitly called for NIST to update SP 800-160 and SP 800-218 for agentic AI, establish minimum engineering requirements covering action authority and tool invocation security, and expand MITRE ATLAS to cover agentic kill-chain tactics, reasoning-layer attacks, and multi-agent lateral movement [7].

The initiative does not yet have specific document numbers or publication timelines for its forthcoming guidelines and research deliverables. The explicit goal of developing interoperable and secure agent deployment guidance represents a significant institutional commitment, but organizations requiring concrete controls today must draw on the parallel workstreams described below rather than waiting for a completed AI Agent Standards Initiative framework.

The NCCoE AI Agent Identity and Authorization Project

On February 5, 2026, the NCCoE published a concept paper titled "Accelerating the Adoption of Software and AI Agent Identity and Authorization," with a public comment deadline of April 2, 2026 [8]. The NCCoE concept paper is notable for its technical specificity, focusing concretely on IAM mechanisms and planning a laboratory demonstration rather than high-level principles—making it the most implementation-oriented deliverable currently available in the federal agentic AI security space.

The concept paper identifies the fundamental problem clearly: existing identity and access management (IAM) frameworks were designed for human users and static software services, neither of which behaves as AI agents do. An AI agent may autonomously access tools, query databases, execute code, and perform operations across multiple systems in a single task, doing so continuously and at a scale and speed that makes traditional per-action human authorization impractical. The paper proposes to address this through four technical focus areas. Identification encompasses distinguishing AI agents from human users and managing the metadata required to scope and bound permissible agent actions. Authorization covers how OAuth 2.0 extensions and policy-based access control mechanisms can be extended to apply to agents as a new class of digital principal. Access delegation addresses how user identities are linked to AI agents in ways that maintain accountability while preventing privilege escalation through delegation chains. Logging and transparency encompasses the mechanisms by which specific AI agent actions are attributed to their non-human entity for audit and forensic purposes.

The NCCoE plans to implement and demonstrate these mechanisms using commercially available technologies in its laboratory environment—a practical proof-of-concept approach that has historically yielded actionable reference architectures for enterprise adoption. Given the NCCoE's track record of producing widely adopted reference architectures, the outcome of this project is a likely candidate to

become a primary federal reference for how enterprises should handle agent authentication, authorization, and audit logging—though COSAiS, NIST IR 8596, or future CISA guidance may also claim significant roles in this space.

COSAiS: SP 800-53 Overlays for AI Agent Systems

The Control Overlays for Securing AI Systems (COSAiS) project, announced by NIST's Computer Security Division in August 2025, is developing SP 800-53 control overlays for five AI use cases [9]. Two of these use cases directly address agentic deployments: "Using AI Agent Systems (Single Agent)" and "Using AI Agent Systems (Multi-Agent)." SP 800-53 is the foundational control catalog for federal information systems and is widely adopted by private sector organizations as a baseline security framework, making these overlays significant not only for federal compliance but for enterprise security programs broadly.

As of March 2026, the COSAiS project has released an annotated outline for its predictive AI use case (published January 8, 2026, with public comment closed February 13, 2026) [10]. Based on the COSAiS project trajectory, the authors anticipate that initial public drafts for the agent system overlays may be released in the coming months, though no specific timeline has been announced by NIST. The completion of these overlays will, for the first time, give organizations a systematic set of SP 800-53 controls specifically tailored to the threat model and deployment characteristics of AI agents—addressing concerns such as least-privilege tool access, agent action containment, multi-agent trust boundaries, and chain-of-custody logging for autonomous operations.

The Empirical Attack Surface

Before the AI Agent Standards Initiative was formally launched, CAISI published empirical research in January 2025 that established the quantitative case for urgency. Working in collaboration with the UK AI Security Institute, NIST researchers tested AI agent hijacking attacks across three priority attack categories: remote code execution through agent tool use, database exfiltration, and automated phishing campaigns conducted via agent communication capabilities [1].

The research, conducted using an enhanced version of the AgentDojo evaluation framework (open-sourced at github.com/usnistgov/agentdojo-inspect), tested defenses against an Anthropic Claude 3.5 Sonnet model deployment. Against baseline prompt-injection defenses, novel attack strategies achieved an 81% success rate, up from an 11% baseline—a 7x improvement in attacker success simply by optimizing the attack prompt. Multi-attempt testing raised the average attack success rate to 80% across the full suite. These findings suggest a structural asymmetry in which the attacker's search space for successful

injections may be substantially larger than the defender's ability to anticipate and block them—at least within current prompt-injection defense paradigms, and across this particular model and evaluation framework.

This empirical work, published by CAISI, is likely informing the NIST AI Agent Standards Initiative's research pillar given the institutional continuity, and the AgentDojo framework is positioned as an ongoing evaluation methodology that can be applied across model providers and agent architectures. The implication for enterprise security teams is that security evaluation of agentic AI deployments should not be treated as a one-time pre-deployment check but as a continuous operational practice.

The Broader Federal Landscape

The NIST workstreams exist within a broader federal policy context that has produced several relevant instruments. NIST IR 8596, the Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile), published in preliminary draft form on December 16, 2025, bridges the CSF 2.0 with the AI RMF and explicitly addresses AI agent use cases as one of its three primary deployment archetypes [11]. While the preliminary draft does not yet contain agentic-specific controls—that work is expected to be incorporated as the COSAiS overlays mature—the Cyber AI Profile establishes the mapping structure under which agent controls will eventually be organized.

CISA has produced two significant documents relevant to agentic deployments. The JCDC AI Cybersecurity Collaboration Playbook, released January 14, 2025, was among the first official US government documents to explicitly name "agentic, copilot, or third-party platforms" in a security reporting context, including them in the information-sharing checklist that organizations are encouraged to complete when reporting AI-related security incidents [12]. In December 2025, CISA joined with the NSA, FBI, and five international partner agencies to publish "Principles for the Secure Integration of AI in Operational Technology," which explicitly defines AI agents as software capable of "autonomous actions" and identifies model drift, hallucinations in control loops, and operator overreliance as agent-specific risks in OT environments [13].

On the procurement and governance side, OMB M-25-21 (April 3, 2025) requires federal agencies to identify and manage "High-Impact AI" systems—defined as AI whose outputs serve as the principal basis for decisions with legal, material, or significant effects [14]. While the memorandum does not enumerate specific technical controls for agentic architectures, many agentic deployments in HR, procurement, logistics, and security operations clearly fall within this definition, triggering requirements for pre-deployment testing, continuous monitoring, and documented human review pathways. OMB M-25-22 established procurement requirements effective for solicitations after October 1, 2025, barring vendors from training publicly available AI on non-public government data and requiring long-term data portability, but it too lacks agentic-specific technical provisions [15].

Identified Gaps

The federal framework for agentic AI security, while rapidly developing, has several significant gaps that directly affect enterprise risk posture. There is no standalone federal security standard specifically for agentic AI systems; the COSAiS overlays under development will fill this function but are not yet published. MITRE ATT&CK for Enterprise and MITRE ATLAS do not currently cover agentic attack patterns—multi-agent lateral movement, tool-use exploitation, delegation chain abuse, and reasoning-layer manipulation represent attacker capabilities that existing threat frameworks do not model [7]. No Federal Acquisition Regulation clause specifically governs the security requirements for AI agent systems acquired by federal agencies; the OMB M-25-22 framework operates through agency-internal procurement procedures rather than codified FAR rules. And the current administration's innovation-first policy orientation creates structural tension with NIST's standards-development timeline: agentic AI is being deployed in federal and enterprise environments at a pace that will significantly precede the availability of the frameworks intended to govern it.

Recommendations

Immediate Actions

Organizations deploying AI agents should evaluate existing deployments against OMB M-25-21's High-Impact AI criteria, even if they are not federal agencies. The High-Impact AI definition—AI whose outputs serve as the principal basis for consequential decisions—is a useful risk filter that captures the agentic deployments most likely to generate compliance exposure and security incidents. Any deployment meeting this standard warrants pre-deployment adversarial testing using evaluation methodologies consistent with the NIST AgentDojo framework, mandatory logging of all agent actions with attribution to specific non-human entities, and documented human review pathways for high-stakes outputs. These controls align with security fundamentals—logging, adversarial testing, and access review documentation—that most mature security programs are already investing in, and will likely be required by forthcoming COSAiS overlays regardless.

Security teams should treat AI agent identity and authorization as a first-order IAM problem, not a future-state consideration. The NCCoE concept paper's four focus areas—identification, authorization, access delegation, and logging—should be applied to existing agent deployments using available IAM tooling even before the NCCoE publishes its full guidance. Specifically, organizations should enforce that

AI agents authenticate as distinct non-human principals (not under user credentials), that tool and API access is scoped via explicit authorization policies rather than inherited from operator permissions, and that delegation chains from user to agent are bounded in scope and duration.

Short-Term Mitigations

While awaiting publication of the COSAiS single-agent and multi-agent overlays, organizations should conduct a gap analysis against NIST AI 100-2 E2025, the updated adversarial machine learning taxonomy published March 24, 2025, which includes the most current NIST taxonomy of agent hijacking attack categories including indirect prompt injection, remote code execution via tool use, and database exfiltration [16]. This document is the closest available published analog to the forthcoming agent-specific controls.

Organizations with AI agents operating in or adjacent to operational technology environments should adopt the CISA/NSA et al. four principles for secure AI integration in OT as a baseline governance framework for those deployments. While developed for OT contexts, the four principles—explainable AI, continuous monitoring, anomaly detection, and enforced human-in-the-loop decision points—are broadly applicable to enterprise AI deployments, with appropriate adaptation to IT-specific risk tolerances.

Security programs should begin tracking NIST's COSAiS annotated outlines and public drafts as they are released, beginning with the predictive AI use case already published, and building internal review capacity to provide substantive feedback during comment periods for the agentic AI overlays. The organizations that engage most substantively with these drafts will have the most influence over the control language that ultimately governs their sector.

Strategic Considerations

The NIST AI Agent Standards Initiative's international engagement pillar—coordinating US positions in ISO/IEC JTC 1 and related bodies—signals that the standards emerging from COSAiS and the NCCoE identity project are intended to become the basis for international consensus. Organizations operating across jurisdictions should anticipate that this federal framework will shape the compliance requirements they face from global regulatory bodies, over a horizon that standards bodies typically measure in 12 to 24 months or more, though international standards convergence timelines are difficult to predict with precision. Organizations that align with NIST's emerging agentic AI controls now will be better positioned to meet compliance requirements from international standards bodies as they converge.

The FDD's March 2026 comment to the NIST RFI articulated a concern worth elevating in enterprise planning: current threat frameworks do not model agentic attack patterns at sufficient granularity for purple-team exercises or detection engineering. Security teams should consider investing in threat modeling using emergent agentic attack taxonomies—including multi-agent lateral movement, orchestrator compromise, and reasoning-layer manipulation—as a parallel track to adopting official guidance, since the official guidance will take time to mature while the threat is present now.

CSA Resource Alignment

The federal framework-building described in this note maps closely onto several active CSA AI Safety Initiative workstreams. The MAESTRO framework for agentic AI threat modeling provides a structured threat taxonomy for multi-agent architectures that complements the NIST COSAiS overlays in development. CSA believes MAESTRO's hierarchical agent layers—from model and data supply chain through orchestration, tool use, and output layers—are broadly consistent with the attack categories NIST has empirically tested and the control areas being defined by the NCCoE identity project, though a formal mapping between MAESTRO and NIST's emerging frameworks has not yet been published. Organizations applying MAESTRO today are working within a threat taxonomy that parallels the federal framework being developed, which may facilitate alignment once the COSAiS overlays are finalized.

The AI Incident Management framework (AICM), as a superset of the Cloud Controls Matrix (CCM) and published by CSA, provides the compliance control structure that organizations using CSA STAR can use to incorporate SP 800-53-aligned controls alongside CCM-mapped requirements within a unified assessment framework. As the COSAiS overlays for single-agent and multi-agent AI systems are published, mapping them to AICM control domains will allow organizations to manage NIST compliance alongside CCM-mapped requirements in a unified assurance program. CSA STAR assessments for organizations deploying AI agents should incorporate the AICM domains most directly implicated by the NCCoE concept paper: identity and access management, audit and accountability, and system and communications protection.

CSA's Zero Trust guidance—specifically the principle of continuous verification and least-privilege access—provides the architectural grounding for the agent identity and authorization controls the NCCoE is developing. AI agents—non-human principals with high-privilege tool access acting on behalf of users—represent a compelling use case for zero trust principles, which were designed to enforce continuous verification and least-privilege access for any principal accessing resources. Ensuring that agent deployments are evaluated through the zero trust lens prior to the publication of federal agent-specific controls provides a defensible, standards-aligned security baseline.

References

[1] National Institute of Standards and Technology, "Technical Blog: Strengthening AI Agent Hijacking Evaluations," NIST CAISI, January 2025. https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

[2] White House, "Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence," January 20, 2025. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

[3] 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

[4] White House, "Ensuring a National Policy Framework for Artificial Intelligence," Executive Order, December 11, 2025. https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/

[5] National Institute of Standards and Technology, "Announcing the AI Agent Standards Initiative: Interoperable and Secure," NIST News, February 17, 2026. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure

[6] National Institute of Standards and Technology, "Request for Information Regarding Security Considerations for Artificial Intelligence Agents," Federal Register, Docket NIST-2025-0035, January 8, 2026. https://www.federalregister.gov/documents/2026/01/08/2026-00206/request-for-information-regarding-security-considerations-for-artificial-intelligence-agents

[7] Foundation for Defense of Democracies, "Security Considerations for AI Agents: FDD Response to NIST-2025-0035," March 9, 2026. https://www.fdd.org/analysis/2026/03/09/regarding-security-considerations-for-artificial-intelligence-agents/

[8] National Cybersecurity Center of Excellence, "Accelerating the Adoption of Software and AI Agent Identity and Authorization," NCCoE Concept Paper, February 5, 2026. https://csrc.nist.gov/pubs/other/2026/02/05/accelerating-the-adoption-of-software-and-ai-agent/ipd

[9] National Institute of Standards and Technology, "NIST Releases Control Overlays for Securing AI Systems Concept Paper," NIST News, August 14, 2025. https://www.nist.gov/news-events/news/2025/08/nist-releases-control-overlays-securing-ai-systems-concept-paper

[10] National Institute of Standards and Technology, "Control Overlays for Securing AI Systems (COSAiS)," CSRC Project Page, 2025–2026. https://csrc.nist.gov/projects/cosais

[11] National Institute of Standards and Technology, "NIST IR 8596 (Initial Preliminary Draft): Cybersecurity Framework Profile for Artificial Intelligence," December 16, 2025. https://csrc.nist.gov/pubs/ir/8596/iprd

[12] Cybersecurity and Infrastructure Security Agency, "JCDC AI Cybersecurity Collaboration Playbook," January 14, 2025. https://www.cisa.gov/resources-tools/resources/ai-cybersecurity-collaboration-playbook

[13] Cybersecurity and Infrastructure Security Agency, National Security Agency, FBI, Australian Signals Directorate's Australian Cyber Security Centre, Canadian Centre for Cyber Security, Germany's BSI, Netherlands' NCSC-NL, and New Zealand's NCSC-NZ, "Principles for the Secure Integration of AI in Operational Technology," December 3, 2025. https://www.cisa.gov/resources-tools/resources/principles-secure-integration-artificial-intelligence-operational-technology

[14] Office of Management and Budget, "M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust," April 3, 2025. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf

[15] Office of Management and Budget, "M-25-22: Driving Efficient Acquisition of Artificial Intelligence in Government," April 3, 2025. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf

[16] National Institute of Standards and Technology, "NIST AI 100-2 E2025: Adversarial Machine Learning — A Taxonomy and Terminology of Attacks and Mitigations," March 24, 2025. https://csrc.nist.gov/pubs/ai/100/2/e2025/final