CSAI Foundation

Cloud Security Alliance AI Safety Initiative

NIST AI Agent Standards: What It Means for Enterprise Security

The Emerging CAISI Framework and Its Implications for Agentic AI Security Programs

Unofficial AI-assisted Research

2026-04-05

Key Takeaways

- NIST's Center for AI Standards and Innovation (CAISI) formally launched the AI Agent Standards Initiative on February 17, 2026 — among the first U.S. government programs dedicated explicitly to security and interoperability standards for agentic AI systems — establishing three strategic pillars: facilitating industry-led standards, fostering open-source protocol development, and advancing research in AI agent security and identity [1].
- The initiative was preceded by a January 8, 2026 Federal Register Request for Information on security considerations for AI agents (docket NIST-2025-0035); the RFI comment period closed March 9, 2026, and respondents including the Foundation for Defense of Democracies explicitly called on NIST to update SP 800-160 and SP 800-218 for agentic AI and to expand MITRE ATLAS to cover multi-agent lateral movement and reasoning-layer attacks [2, 3].
- NIST's National Cybersecurity Center of Excellence (NCCoE) published a companion concept paper on February 5, 2026 — "Accelerating the Adoption of Software and AI Agent Identity and Authorization" — that directly addresses the enterprise security gap where AI agents are commonly treated as generic service accounts with no dedicated identity, authorization, or accountability controls [4].
- The COSAiS project, announced in mid-2025, is developing SP 800-53 control overlays for five AI deployment categories including single-agent and multi-agent systems — providing enterprises with a structured path to map existing NIST controls to agentic deployments before final guidance is published [5].
- NIST IR 8596, a preliminary draft Cybersecurity Framework Profile for AI published December 16, 2025, provides the most direct bridge for organizations running CSF 2.0 programs to incorporate AI-specific risk categories without rebuilding their security program architecture [6].
- Enterprises should not wait for final NIST publications: the draft guidance, the NCCoE concept paper, and the updated AI 100-2 adversarial taxonomy collectively define the expected direction of the final standards and represent the leading edge of what regulators and auditors will reference in 2026 assessments [7].

Background

The security risk posed by AI agents is qualitatively different from prior enterprise AI deployments. Earlier-generation chatbots and recommendation systems without tool access operate within bounded, largely read-only contexts; their failure modes are generally limited to incorrect outputs. Agentic AI systems — autonomous programs capable of planning multi-step tasks, invoking tools, executing code, delegating to sub-agents, and persisting state across sessions — interact with enterprise infrastructure in ways that create novel attack surfaces, blur accountability boundaries, and generate entirely new categories of privilege escalation. An agent authorized to draft and send email on behalf of a user can be manipulated into exfiltrating sensitive data through that same channel. An agent authorized to query a database can be directed, via a poisoned tool response, to alter it. These are not theoretical scenarios; NIST's own empirical red-team research documented an 81% attack success rate against AI agent systems using novel adversarial strategies in controlled exercises [12]. The threat is operational, not hypothetical.

The regulatory and standards landscape has lagged this operational reality. Existing NIST frameworks — the AI Risk Management Framework (AI RMF 1.0) [11], the Cybersecurity Framework (CSF 2.0), and the SP 800-53 control catalog — were designed before the emergence of production-scale agentic deployments. They provide useful foundations but contain systematic gaps in the control families most critical for agentic systems: AC (Access Control), IA (Identification and Authentication), AU (Audit and Accountability), and SR (Supply Chain Risk Management) [5]. The SP 800-53 catalog has well-developed controls for human and system identity, but lacks purpose-built controls for distinguishing an AI agent from a human operator, scoping agent permissions to a defined task context, or linking agent actions to a non-human principal for forensic attribution. Existing IA and AC controls provide partial coverage, but no SP 800-53 controls were designed with AI agent principal distinctions in mind.

NIST's response to this gap is a coordinated cluster of initiatives launched between mid-2025 and early 2026. The AI Agent Standards Initiative, the NCCoE agent identity concept paper, the COSAiS overlay project, and the IR 8596 Cyber AI Profile collectively represent a substantial realignment of the U.S. federal AI security standards posture since the original AI RMF release in January 2023. Enterprises that have built security programs around NIST frameworks need to understand both the current state of these initiatives and the practical steps available to address agentic security gaps before final guidance is published — because enforcement and audit cycles will not pause for the publication schedule.

Security Analysis

The AI Agent Standards Initiative: Strategic Architecture

The CAISI initiative, announced February 17, 2026, is organized around three interdependent pillars that reflect NIST's recognition that agent security is simultaneously a technical, ecosystem, and geopolitical challenge [1]. The first pillar — facilitating industry-led standards — focuses on technical convenings, gap analyses, and voluntary guidelines, with explicit intent to strengthen U.S. stakeholder participation in international standards-setting through ISO/IEC JTC 1. The CAISI announcement directly references international competition in AI standards as a motivating factor, and the Foundation for Defense of Democracies' March 2026 RFI response reinforced that framing, urging NIST to accelerate deliverables to maintain U.S. influence over the emerging international agentic AI standards landscape [3].

The second pillar addresses open-source protocol interoperability — a pointed acknowledgment that the emerging ecosystem of agentic frameworks (including Model Context Protocol, agent-to-agent communication protocols, and tool invocation standards) is developing faster than any single vendor or standards body can track. NIST's approach here is explicitly facilitative rather than prescriptive: the agency will identify barriers to secure interoperability and co-invest with the National Science Foundation through its Pathways to Enable Secure Open-Source Ecosystems program, rather than attempting to standardize specific protocols that may not yet be stable [1]. The third pillar — foundational research in agent authentication and identity infrastructure — is the most technical and the most directly actionable for enterprise security programs, as it feeds directly into the NCCoE concept paper work described below.

Key planned deliverables include an AI Agent Interoperability Profile (Q4 2026), which will represent the initiative's first comprehensive normative output. Sector-specific listening sessions in healthcare, finance, and education have already occurred as of the publication date of this note. The initiative's RFI — Federal Register docket NIST-2025-0035 — received responses from OpenID Foundation, commercial AI providers, and national security-focused organizations through its March 9, 2026 deadline. These responses are expected to shape the technical direction of forthcoming draft publications [2].

The Agent Identity and Authorization Gap

The NCCoE concept paper published February 5, 2026 is arguably the most immediately actionable output of NIST's agentic AI standards cluster for enterprise security teams without an existing FedRAMP or SP 800-53 baseline, given its concrete focus on identity architecture gaps [4]. It frames the core security problem with precision: current enterprise identity infrastructure was designed for human users

and conventional software services. AI agents do not fit cleanly into either category. They are not human users — they cannot exercise judgment about appropriate scope — but they are also not static services, because their behavior adapts dynamically based on goals, context, and tool outputs. Treating an AI agent as a service account grants it indefinite standing permissions with no mechanism for scoping those permissions to a specific task or time window. Treating it as a delegated user collapses accountability by merging agent actions into the identity record of the human principal it serves.

The concept paper proposes four technical focus areas to close this gap. The first is identification: developing mechanisms to distinguish AI agents from human users in authentication flows, and to capture metadata defining the permissible range of agent actions as part of the agent's identity record. The second is authorization: extending OAuth 2.0 and policy-based access control (PBAC) frameworks with agent-specific permission scoping, so that a user delegating email management to an agent does not inadvertently grant that agent access to financial systems. The third is access delegation: designing protocols that link user identities to AI agents in a way that preserves accountability — the agent acts, but the action is attributable to a named principal — without conflating user identity and agent identity. The fourth is logging and transparency: ensuring that agent actions generate audit records attributed to non-human entities, enabling forensic investigation of agent behavior without requiring manual reconstruction from first principles [4].

The NCCoE held its concept paper comment period open through April 2, 2026. The eventual project deliverable will be a practice guide providing reference implementations for each focus area. Enterprises that have previously engaged with NCCoE practice guides on topics such as zero trust architecture (SP 1800-35) will recognize the format: vendor-agnostic reference architectures, implemented in testbed environments, with mapping to NIST SP 800-53 controls. Based on NCCoE's established pattern, the eventual practice guide is expected to follow the same vendor-agnostic reference architecture format.

SP 800-53 Control Overlays: Filling the Gaps Today

The COSAiS project, whose concept paper was published in mid-2025, takes a different approach to the same problem: rather than waiting for new controls to be developed, it maps existing SP 800-53 controls as overlays targeted to specific AI deployment categories [5]. For enterprise security teams already operating SP 800-53-based control environments — FedRAMP authorized services, DoD IL environments, NIST-aligned security programs — this approach offers a path to address AI-specific gaps using existing authorization and audit machinery.

Two of COSAiS's five use-case categories are directly relevant to agentic deployments: "Using AI Agent Systems — Single Agent" (autonomous systems with minimal human oversight) and "Using AI Agent Systems — Multi-Agent" (cooperative autonomous agents with limited human supervision). The concept paper's gap analysis identified the most severe deficiencies in four SP 800-53 control families: AC

(Access Control), IA (Identification and Authentication), AU (Audit and Accountability), and SR (Supply Chain Risk Management). These map precisely to the threats that make agentic AI systems attractive targets — privilege escalation through unbounded permissions, accountability gaps when agents act across system boundaries, auditability failures when agent actions are attributed to human identities, and supply chain risks introduced by third-party tool integrations and model providers.

The COSAiS overlays are still in development; the project page lists several resources as pending, and no final overlay documents have been published as of the date of this note. However, the concept paper's framing is sufficiently detailed that security architects can begin mapping their existing SP 800-53 control implementations against the identified gap areas now. Organizations waiting for the final overlays before beginning this analysis will find themselves behind the curve when auditors begin requesting AI-specific control evidence in 2026 FedRAMP and DoD assessments.

The Adversarial Threat Landscape: AI 100-2 and MITRE ATLAS

NIST's standards initiatives do not exist in isolation from the threat landscape they are designed to address. The March 2025 update to AI 100-2 — the adversarial machine learning taxonomy — substantially expanded the catalog of agentic-AI-specific attacks that the new control overlays and identity guidance are intended to mitigate [7]. The updated taxonomy introduces attack categories with no traditional cybersecurity analog, including indirect prompt injection via tool invocation (adversary-controlled content in tool outputs redirecting agent goals), RAG-targeted attacks that manipulate retrieval-augmented generation contexts used by agent reasoning, and multi-step agentic attacks that chain tool misuse across multiple agent invocations to achieve objectives — such as database exfiltration or code execution — that would be prevented by any single control in isolation.

MITRE ATLAS has tracked this threat landscape in parallel, expanding from a primarily academic ML security taxonomy to a detailed agentic adversary framework through 2025–2026. The November 2025 v5.1.0 release substantially expanded the framework's coverage of agentic techniques, mitigations, and case studies [8]. The February 2026 v5.4.0 update added techniques specifically targeting the agentic tool ecosystem, including "Publish Poisoned AI Agent Tool" — directly relevant to the MCP and A2A protocol attack surface — and "Escape to Host," which catalogs how agent systems with code execution capabilities can be leveraged to break out of their intended operational context. The Foundation for Defense of Democracies' March 2026 NIST RFI response explicitly called for formal coordination between NIST and MITRE to incorporate these agentic kill-chain tactics into NIST guidance, suggesting that alignment between the two frameworks is an active policy conversation [3].

OWASP's Top 10 for Agentic Applications 2026, published December 10, 2025 and community-reviewed by over 100 security researchers, provides the most actionable practitioner-level threat model currently available [9]. The list's framing principle — "Least Agency," the agentic equivalent of least privilege —

maps directly to the NCCoE concept paper's authorization focus areas and provides a concrete threat-modeling vocabulary for enterprise security teams evaluating agentic deployments before NIST final guidance is available.

Recommendations

Immediate Actions

Enterprise security teams should treat the February 2026 NCCoE concept paper on AI Agent Identity and Authorization as current best-practice guidance, not a preview of future requirements [4]. The four focus areas it identifies — identification, authorization, access delegation, and logging — define the minimum security architecture for any production agentic deployment. Organizations should audit existing AI agent deployments against these criteria immediately: does each agent have a distinct identity separate from the human principal it serves? Are agent permissions scoped to the minimum required for each task? Do audit logs attribute agent actions to non-human entities in a way that supports forensic investigation? If the answer to any of these questions is no, the deployment presents a significant accountability risk that warrants prioritized remediation before final NIST guidance is published.

Security architects operating in SP 800-53 environments should begin the COSAiS gap analysis now, focusing on the AC, IA, AU, and SR control families identified in the concept paper [5]. The forthcoming overlays will provide normative mapping, but the control gaps are already described with sufficient precision in the concept paper to support preliminary assessment and remediation planning. Organizations that complete this analysis before the final overlays are published will be positioned to treat the overlay release as a validation exercise rather than the start of a remediation cycle.

The OWASP Top 10 for Agentic Applications 2026 should be incorporated immediately into threat modeling processes for any new agentic AI system under development or procurement evaluation [9]. The Least Agency principle — minimum autonomy, minimum tool access, minimum credential scope — should be a design requirement, not a security review comment. Agents that operate with standing access to enterprise systems rather than task-scoped credentials represent an elevated privilege escalation risk and should be treated as a priority remediation item.

Short-Term Mitigations

Organizations should review their enterprise identity provider (IdP) configurations to evaluate whether existing service account management capabilities can be adapted to enforce key aspects of the NCCoE's agent identity model before purpose-built agent identity controls are available. OAuth 2.0 PBAC extensions and token scoping mechanisms available in current IdP platforms can approximate agent-specific authorization constraints, and applying them now creates an audit trail that can be mapped to forthcoming NIST controls as they are published. Organizations using major cloud identity platforms — Entra ID, Okta, Google Workspace — should work with their vendors to understand native AI agent identity support, as major identity platform vendors are likely to develop agent-specific identity primitives given customer demand and the direction of NIST's emerging standards.

Audit logging for AI agent systems should be reviewed to ensure agent actions are recorded in a format that supports the forensic attribution requirements described in the NCCoE concept paper. At minimum, each agent invocation should be logged with: the initiating human principal's identity; the agent's identifier; the specific tools or APIs invoked; the data accessed or modified; and the session context. Organizations that have deployed AI agents as extensions of human user sessions, where agent actions blend into the human user's audit record, should treat this as a high-priority remediation item — this pattern directly prevents effective incident investigation.

Strategic Considerations

The NIST AI Agent Standards Initiative's explicit focus on ISO/IEC JTC 1 engagement signals that agentic AI security standards will increasingly be developed and evaluated in international forums where ISO/IEC 42001 — the AI Management System standard — will serve as the foundational governance reference [10]. Organizations that have not yet mapped their AI security programs to ISO/IEC 42001 should begin that assessment, not because 42001 directly addresses agentic security, but because the emerging international standards ecosystem is expected to build on 42001 as a governance foundation — particularly for international audit and certification schemes. Supply chain partners, enterprise customers, and regulators outside the United States are increasingly requiring 42001 alignment alongside AI RMF mapping.

The coordinated NIST work cluster — AI Agent Standards Initiative, NCCoE agent identity project, COSAiS overlays, IR 8596 Cyber AI Profile — represents the pre-publication state of what may become baseline requirements for federal contractors and FedRAMP-authorized services within the next several years, based on prior patterns of NIST-to-FedRAMP adoption. Commercial enterprises in regulated industries (financial services, healthcare, critical infrastructure) should anticipate that state and sector

regulators will incorporate these frameworks into AI governance guidance on a similar timeline. Organizations that treat NIST's draft guidance as optional reading risk facing a compliance gap when final publications align with examination cycles.

CSA Resource Alignment

The security challenges addressed by NIST's AI agent standards cluster map directly to foundational CSA frameworks that provide complementary guidance for enterprise practitioners. The MAESTRO framework for agentic AI threat modeling addresses the same multi-layer attack surface that motivates NIST's AI 100-2 taxonomy updates and the MITRE ATLAS agentic technique expansions — specifically the interaction between model layer attacks (prompt injection, goal hijacking) and infrastructure layer attacks (tool misuse, privilege escalation) that make agentic deployments uniquely difficult to defend with controls designed for static systems. Organizations implementing the NIST COSAiS overlays for single-agent and multi-agent systems should use MAESTRO as the threat model that informs control selection, ensuring that the SP 800-53 control gaps identified by NIST are mapped to the specific threat scenarios most likely to affect the organization's deployment architecture.

The CSA AI Organizational Responsibilities framework provides the governance accountability structure within which NIST's agent identity and authorization requirements sit. The NCCoE concept paper's focus on access delegation — linking agent actions to named human principals — is an implementation-layer expression of the organizational accountability principles that CSA has articulated at the governance layer. Enterprises that have adopted the CSA AI Organizational Responsibilities framework already have the governance foundation to implement the NCCoE's technical identity model; the question is whether that governance accountability has been extended to cover agentic deployments specifically.

The CSA Cloud Controls Matrix (CCM), and by extension the AI Controls Matrix (AICM) — which extends CCM with AI-specific controls and is the preferred reference for AI deployments — provides a cross-framework mapping surface that enterprises can use to align NIST's emerging agentic AI controls with their existing compliance posture. As the COSAiS project publishes its SP 800-53 overlays for AI agent systems, AICM users should expect those overlays to be incorporated into AICM mapping tables, enabling organizations to satisfy both U.S. federal requirements and broader cloud security compliance obligations through a unified control set. CSA's STAR program for AI security attestation offers a mechanism for organizations and vendors to demonstrate conformance with these emerging standards before formal certification schemes are available from NIST or accreditation bodies.

References

[1] NIST CAISI. "Announcing the AI Agent Standards Initiative for Interoperable and Secure AI Agents." NIST, February 17, 2026.

[2] Federal Register. "Request for Information Regarding Security Considerations for Artificial Intelligence Agents." Docket NIST-2025-0035, January 8, 2026.

[3] Foundation for Defense of Democracies. "Regarding Security Considerations for Artificial Intelligence Agents." FDD, March 9, 2026.

[4] NCCoE. "Accelerating the Adoption of Software and AI Agent Identity and Authorization." NIST NCCoE Concept Paper, February 5, 2026.

[5] NIST CSRC. "Control Overlays for Securing AI Systems (COSAiS)." NIST Project Page, announced mid-2025.

[6] NIST. "NIST IR 8596 (Initial Preliminary Draft): Cybersecurity Framework Profile for Artificial Intelligence." NIST, December 16, 2025.

[7] NIST. "NIST AI 100-2 E2025: Adversarial Machine Learning — A Taxonomy and Terminology of Attacks and Mitigations." NIST, March 2025.

[8] MITRE. "MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems." MITRE, v5.4.0 February 2026.

[9] OWASP GenAI. "OWASP Top 10 for Agentic Applications 2026." OWASP, December 10, 2025.

[10] ISO. "ISO/IEC 42001:2023 — Information Technology — Artificial Intelligence — Management System." International Organization for Standardization, 2023.

[11] NIST. "AI Risk Management Framework (AI RMF 1.0)." NIST AI 100-1, January 26, 2023.

[12] NIST CAISI. "Insights into AI Agent Security from a Large-Scale Red-Teaming Competition." NIST CAISI Research Blog, January 2025.

