NIST AI Agent Standards: Navigating the Federal Governance Gap

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-28

Key Takeaways

- On February 17, 2026, NIST's Center for AI Standards and Innovation (CAISI) launched the AI Agent Standards Initiative — the first U.S. federal program specifically targeting the security, interoperability, and identity challenges of autonomous AI agents [1].
- The initiative arrived against a backdrop of significant federal policy turbulence: the Biden administration's October 2023 AI Executive Order was revoked in January 2025, creating a period during which voluntary NIST frameworks became the primary governance instruments available to most organizations [2][3].
- NIST's simultaneous work — a published RFI on AI agent security, a National Cybersecurity Center of Excellence (NCCoE) concept paper on agent identity and authorization, and a preliminary draft Cyber AI Profile (NISTIR 8596) — constitutes the federal government's most extensive published effort to date to define a governance vocabulary for agentic AI systems, even as none of these documents carries regulatory force [4][5][6].
- Survey data from early 2026 illustrates the governance gap from two angles: Microsoft found that over 80% of Fortune 500 companies are actively deploying AI agents [7], while a separate Gravitee survey of enterprise teams found only 14.4% had full security and IT approval for their agentic deployments [8] — findings from different sample populations that nonetheless converge on the same concern.
- The NCCoE concept paper on AI agent identity and authorization accepts public comment through April 2, 2026, and CAISI is scheduling sector listening sessions across healthcare, finance, and education through April 2026 — creating an immediate window for industry input into the baseline governance assumptions that will shape the AI Agent Interoperability Profile expected in late 2026 [1][5].

Background

The February 2026 NIST AI Agent Standards Initiative

On February 17, 2026, NIST announced the launch of its AI Agent Standards Initiative through CAISI — the Center for AI Standards and Innovation — framing its mission as ensuring that autonomous AI systems "can function securely on behalf of their users, and can interoperate smoothly" [1]. The initiative is organized around three strategic pillars. The first focuses on facilitating industry-led standards development and advancing U.S. leadership in international standards bodies, in coordination with the National Science Foundation. The second supports community-led open-source protocol development through NSF's Pathways program. The third pursues technical research into AI agent security, authentication, identity infrastructure, and security evaluation methodologies [1].

The February announcement did not emerge in isolation. It followed two months of CAISI preparatory activity that gives a clearer picture of where NIST sees the technical risk. On January 8, 2026, NIST published a Request for Information on AI agent security (Docket NIST-2025-0035) in the Federal Register, with a comment period that closed March 9, 2026 [4]. The RFI identified three threat categories that characterize the distinct risk profile of agentic systems: adversarial data interactions including indirect prompt injection attacks capable of hijacking agent behavior; compromised models subject to data poisoning or other corruption; and misaligned objectives, in which agents pursue unintended goals through specification gaming even without adversarial inputs. These categories represent NIST's emerging analytical vocabulary for agentic risk, and they signal that the agency's forthcoming guidance will address attack vectors largely absent from its prior AI documents.

On February 5, 2026, NIST's NCCoE published a concept paper titled "Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization," with a public comment period running through April 2, 2026 [5]. The document proposes a demonstration project addressing four specific technical challenges: identifying AI agents as distinct from human users and managing their associated metadata; implementing authorization controls governing what AI agents are permitted to do; establishing auditing and non-repudiation mechanisms that connect specific agent actions to an auditable accountability record; and preventing and mitigating prompt injection techniques capable of subverting agent behavior [5]. The prominence given to prompt injection as a named challenge reflects an acknowledgment that existing authentication and authorization standards were not designed with agentic threat models in mind.

Federal Policy Instability and the Governance Vacuum

The NIST initiative is taking shape within a federal AI governance landscape that has undergone substantial disruption. The Biden administration's Executive Order 14110, issued October 30, 2023, established an extensive set of requirements for AI safety evaluation, federal agency AI governance, and developer transparency obligations [2]. That order was revoked on January 20, 2025, within hours of the Trump administration taking office; Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence," was issued three days later, on January 23, 2025, directing agencies to rescind or revise any policies that conflicted with a pro-innovation framing [3]. In practice, this transition substantially reduced formal federal requirements for most agencies, with voluntary NIST frameworks becoming — in the absence of binding successors — the primary available governance vocabulary for organizations seeking AI policy guidance [9].

Subsequent executive action has moved toward a lighter-touch federal posture. The Office of Management and Budget's memorandum M-25-21, "Accelerating Federal Use of AI through Innovation, Governance, and Public Trust," issued April 3, 2025, required agencies to appoint Chief AI Officers and established risk management obligations for "high-impact AI" — systems whose outputs serve as "a principal basis for decisions or actions with legal, material, binding, or significant effect" [9]. The memorandum acknowledged agentic workflows as an emerging category, describing AI-powered systems that "chain tasks together," but did not establish specific requirements for them. The White House AI Action Plan, released July 23, 2025, directed NIST to revise the AI RMF 1.0 to remove references to misinformation, diversity and inclusion, and climate change — a content-specific directive to a technical standards body with few recent precedents in U.S. standards governance [10].

The December 11, 2025 Executive Order "Ensuring a National Policy Framework for Artificial Intelligence" added a further dimension by directing the Department of Justice to establish an AI Litigation Task Force to challenge state AI laws deemed to burden interstate commerce or otherwise conflict with federal policy [11]. This created a federal-state tension in which the federal government simultaneously pursues minimal binding obligations for AI at the national level while seeking to challenge more prescriptive state regulatory activity through federal litigation [11]. For organizations operating across jurisdictions, this dynamic creates compliance uncertainty without providing the clarity of a binding federal standard.

No binding federal AI statute has been enacted. The 119th Congress has introduced multiple bills addressing AI governance, including the Artificial Intelligence Risk Evaluation Act of 2025 (S.2938), which defines advanced AI as systems capable of autonomous operation in open-ended environments that could modify their own functions to circumvent human control [12], and the Future of Artificial

Intelligence Innovation Act of 2026 (S.3952) [18]. Neither has advanced to a floor vote. In this environment, NIST's voluntary frameworks remain the primary governance vocabulary for most organizations.

Security Analysis

The Technical Gaps in Existing Frameworks

The NIST AI Agent Standards Initiative is responding to a threat surface that existing frameworks have not yet fully characterized. NIST AI 100-2 E2025, published March 24, 2025, updated the adversarial machine learning taxonomy to incorporate generative AI and introduced "explicit guidance on securing AI supply chains, dealing with risks posed by autonomous AI agents, and securing enterprise-grade GenAI integrations" — the first time autonomous AI agent deployments were addressed as a distinct category in NIST's published AI security taxonomy [13]. But AI 100-2 is a taxonomy document, not a controls framework. The preliminary draft NISTIR 8596 Cyber AI Profile, published December 16, 2025, is intended to bridge that gap, but the draft does not yet fully characterize the controls needed for multi-agent systems in which multiple agents coordinate and take autonomous action [6]. NIST's planned SP 800-53 Control Overlays for Securing AI Systems (COSAiS) project, with a concept paper published in August 2025, is expected to introduce controls for single and multi-agent AI systems, but a public draft had not been published as of the date of this note [14].

The identity and authorization problem receives the most immediate attention, and for good reason. Existing authentication infrastructure was designed to manage human identities and service accounts with well-understood privilege models. AI agents introduce a new identity category: a non-human actor that operates with delegated human authority, dynamically acquires tool-use capabilities, spawns sub-agents that inherit or expand that authority, and persists across sessions in ways that create authorization state the NCCoE concept paper identifies as a novel governance challenge [5]. The NCCoE concept paper acknowledges that most enterprises currently rely on shared API keys to authenticate agents — a control model that provides no meaningful accountability for individual agent actions and no basis for least-privilege enforcement at the agent level [5]. The OWASP Top 10 for Agentic Applications 2026, released December 10, 2025, identifies identity and privilege abuse (ASI03) and insecure inter-agent communication (ASI07) among the leading risks, with spoofed inter-agent messages capable of exploiting trust relationships between orchestrator and sub-agent components [15].

The Enterprise Governance Gap

Survey data published in the first quarter of 2026 paints a concerning picture of the gap between agentic AI deployment velocity and governance maturity. Microsoft's Cyber Pulse report of February 10, 2026, found that 80% of Fortune 500 companies are actively deploying AI agents [7]. A Gravitee survey found that 81% of enterprise teams had moved past planning into active testing or production for agentic systems, yet only 14.4% had full security and IT approval for those deployments — a finding from a different sample population, but one that directionally reinforces the Microsoft data [8]. HiddenLayer's 2026 AI Threat Landscape Report, published March 18, 2026, surveyed 250 IT and security leaders and found that one in eight reported AI breaches now involves agentic systems, that 31% of organizations cannot confirm whether they experienced an AI security breach in the past year, and that 53% admitted withholding breach reports despite 85% supporting mandatory disclosure requirements [16]. Perhaps most operationally significant, 60% of organizations reported that they cannot terminate a misbehaving agent — a capability that traditional data loss prevention and endpoint security tools were not designed to provide [17].

The governance gap has a structural dimension that voluntary frameworks alone cannot close. NIST AI RMF and ISO 42001 provide organizational governance structures — risk committees, documentation requirements, accountability roles — but neither addresses the specific technical controls that security teams require for agentic deployments: tool call parameter validation, prompt injection logging, containment testing for multi-agent systems, or behavioral drift detection for long-running agents. Industry response to NIST's January 2026 RFI reflected this tension; TechNet's comment letter urged NIST to keep agentic AI standards flexible and voluntary, arguing that prescriptive controls would constrain innovation at a critical development phase [19]. This voluntary-versus-mandatory tension mirrors the broader federal posture and suggests that binding technical requirements for agentic AI appear unlikely in the near term given current political conditions, with no pending federal legislation having advanced to a floor vote [12].

Recommendations

Immediate Actions

Organizations deploying or evaluating agentic AI should engage directly with NIST's ongoing public participation processes. The NCCoE concept paper on AI agent identity and authorization accepts comments through April 2, 2026, and organizations with deployment experience in agent IAM, OAuth integration for non-human actors, or prompt injection mitigation are particularly well-positioned to

contribute input that will shape the eventual NCCoE demonstration project [5]. In the CSA AI Safety Initiative's assessment, industry input during concept paper comment periods provides greater latitude to influence scope and framing than do comment periods on complete draft documents — making early engagement particularly consequential.

Enterprises should also conduct an immediate inventory of current agent identity practices. For each deployed agent or agent system, organizations should document the identity model in use (dedicated service account, shared API key, delegated OAuth token, or other), the effective privilege scope including all accessible tools, APIs, and data stores, and the authorization chain from the human or system principal that delegated authority to the agent. Organizations that rely on shared API keys for agent authentication should prioritize migration to per-agent identity with scoped credentials and expiration controls.

Short-Term Mitigations

Over the next three to six months, organizations should implement a formal agent authorization model that explicitly treats AI agents as a distinct identity class rather than as generic software service accounts. This entails adapting existing OAuth 2.0 or similar authorization frameworks to issue agent-specific tokens scoped to minimum required capabilities, with delegation relationships recorded in an auditable identity store. The NCCoE concept paper's focus on access delegation — linking agent identity to the human principal that authorized it — provides a useful design target even before formal NIST guidance is published [5].

Organizations should adopt the OWASP Top 10 for Agentic Applications 2026 as an interim technical controls checklist, with particular attention to agent goal hijacking via prompt injection (ASI01), tool misuse (ASI02), identity and privilege abuse (ASI03), and insecure inter-agent communication (ASI07) [15]. These categories represent the risks the OWASP working group identified as most prevalent in early agentic deployments and provide a vendor-neutral framework for gap assessment against which emerging NIST guidance can later be mapped.

Runtime logging should be extended to capture agent actions at the tool-call level, not merely at the session or request level. This includes recording the inputs and outputs of each tool invocation, the agent identity and authorization token in use at the time of invocation, and any sub-agent delegation events. This logging posture is explicitly anticipated by the NCCoE concept paper and will position organizations favorably for compliance with forthcoming formal guidance [5].

Strategic Considerations

At the strategic level, the current voluntary framework environment represents an opportunity for organizations to influence the standards that will eventually govern their deployments. Industry participation in NIST sector listening sessions — CAISI is scheduling sessions across healthcare, finance, and education through April 2026 — offers direct input into baseline assumptions that will shape the AI Agent Interoperability Profile expected in late 2026 [1]. Organizations with mature agentic AI programs should consider submitting case studies, contributing to OWASP working groups, and engaging with CSA's AI Safety Initiative to ensure that practical deployment experience informs rather than simply reacts to the governance frameworks under development.

Organizations should also monitor the federal-state dynamic carefully. The December 2025 executive order directing DOJ to challenge state AI laws creates uncertainty for compliance programs built around emerging state requirements, but does not eliminate state enforcement activity in the near term [11]. Compliance architectures should be designed for governance pluralism — capable of satisfying overlapping requirements from federal guidance, state law, and sector regulators such as the OCC, FRB, or HHS Office for Civil Rights — rather than assuming a unified federal standard will consolidate the landscape in the near term.

CSA Resource Alignment

This note connects to several active CSA AI Safety Initiative resources. The CSA AI Controls Matrix (AICM), published July 2025, is the most directly applicable: its 243 controls across 18 domains include dedicated coverage for autonomous agent governance, access delegation, and runtime behavioral monitoring that complements the identity and authorization gaps NIST's NCCoE concept paper explicitly targets. Organizations mapping to NIST AI RMF should use AICM as the control specification layer for the agentic extensions the RMF currently lacks.

The CSA MAESTRO threat modeling framework provides structured guidance for the threat categories that NIST's January 2026 RFI identified — adversarial data interactions, model compromise, and misaligned objectives — and maps each to mitigation strategies applicable to multi-agent architectures. MAESTRO's integration with the AICM provides a direct path from threat enumeration to control selection for organizations building agentic governance programs.

The CSA whitepaper "NIST AI Risk Management Framework: Agentic Profile" (published March 27, 2026) provides a detailed gap analysis of RMF 1.0 for agentic deployments and proposes a structured set of extensions across the four RMF functions — GOVERN, MAP, MEASURE, and MANAGE — that

organizations can implement ahead of NIST's formal guidance. That document addresses autonomy tier classification, tool-use risk modeling, runtime behavioral metrics, and delegation chain accountability, all of which align with the CAISI initiative's announced focus areas. Organizations should read the two documents together: this research note describes the federal policy and standards context, while the Agentic Profile document provides the technical framework for organizational implementation.

The AAGATE reference architecture, published by CSA in December 2025, translates these RMF and AICM principles into a Kubernetes-native runtime governance overlay, providing engineering teams with implementation patterns for the agent identity, logging, and containment controls described in both this note and the emerging NIST guidance [20].

References

[1] NIST Center for AI Standards and Innovation, "Announcing the AI Agent Standards Initiative," NIST News, February 17, 2026. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure

[2] White House, "Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," October 30, 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[3] White House, "Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence," January 23, 2025. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

[4] NIST CAISI, "CAISI Issues Request for Information About Securing AI Agent Systems," NIST News, January 12, 2026 (Federal Register Docket NIST-2025-0035, published January 8, 2026). https://www.nist.gov/news-events/news/2026/01/caisi-issues-request-information-about-securing-ai-agent-systems

[5] NIST NCCoE, "Accelerating the Adoption of Software and Artificial Intelligence 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

[6] NIST, "Draft NIST Guidelines Rethink Cybersecurity for the AI Era," NIST News, December 16, 2025 (NISTIR 8596, Preliminary Draft (IPRD)). https://www.nist.gov/news-events/news/2025/12/draft-nist-guidelines-rethink-cybersecurity-ai-era

[7] Microsoft Security, "80% of Fortune 500 Use Active AI Agents: Observability, Governance, and Security Shape the New Frontier," Microsoft Security Blog, February 10, 2026. https://www.microsoft.com/en-us/security/blog/2026/02/10/80-of-fortune-500-use-active-ai-agents-observability-governance-and-security-shape-the-new-frontier/

[8] Gravitee, "State of AI Agent Security 2026: When Adoption Outpaces Control," Gravitee Research, 2026. https://www.gravitee.io/blog/state-of-ai-agent-security-2026-report-when-adoption-outpaces-control

[9] 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

[10] White House, "America's AI Action Plan: Winning the Race," July 23, 2025. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[11] 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/

[12] 119th Congress, S.2938 "Artificial Intelligence Risk Evaluation Act of 2025." https://www.congress.gov/bill/119th-congress/senate-bill/2938/text

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

[14] NIST, "NIST Releases Control Overlays for Securing AI Systems Concept Paper," NIST News, August 14, 2025 (COSAiS Project). https://www.nist.gov/news-events/news/2025/08/nist-releases-control-overlays-securing-ai-systems-concept-paper

[15] OWASP GenAI Security Project, "OWASP Top 10 for Agentic Applications 2026," December 10, 2025. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

[16] HiddenLayer, "2026 AI Threat Landscape Report," March 18, 2026. https://www.hiddenlayer.com/news/hiddenlayer-releases-the-2026-ai-threat-landscape-report-spotlighting-the-rise-of-agentic-ai-and-the-expanding-attack-surface-of-autonomous-systems

[17] Kiteworks, "2026 Data Security and Compliance Risk Forecast Report," Kiteworks Research, 2026.

[18] 119th Congress, S.3952 "Future of Artificial Intelligence Innovation Act of 2026," introduced February 26, 2026. https://www.congress.gov/bill/119th-congress/senate-bill/3952/all-info

[19] TechNet, "Public Comment on NIST-2025-0035 (AI Agent Security RFI)," March 2026. [Available via Federal Register Docket NIST-2025-0035 public comment record]

[20] Cloud Security Alliance, "AAGATE: Agentic AI Governance Architecture for Trusted Execution," CSA Research Publication, December 2025.