CSAI Foundation

Cloud Security Alliance AI Safety Initiative

Agentic AI's Governance Vacuum: NIST Standards and the Gap

Why Enterprises Cannot Wait for Standards That Won't Arrive Until 2027

Unofficial AI-assisted Research

2026-04-12

Key Takeaways

- NIST's Center for AI Standards and Innovation (CAISI) launched the **AI Agent Standards Initiative** in February 2026, establishing a U.S. government program specifically dedicated to agentic AI governance — but its most specific technical guidance, the SP 800-53 control overlays for single-agent and multi-agent systems (COSAiS), is not expected until late 2026 to 2027 [1][2].
- The standards gap is not theoretical: NIST's own red-team research demonstrated that novel attack techniques against AI agents achieve an **81% task-hijacking success rate** against frontier model deployments, up from 11% for baseline defenses [3].
- Enterprise deployments are outpacing governance. Only **23% of organizations** have a formal, enterprise-wide strategy for AI agent identity management, while **44%** rely on static API keys — a credential pattern that, as typically deployed, lacks rotation schedules and usage monitoring — to authenticate autonomous agents [4].
- The NCCoE's February 2026 concept paper on AI agent identity and authorization closed its public comment period on April 2, 2026, with a full demonstration project yet to be scoped [5].
- Industry projections suggest non-human and agentic identities could reach tens of billions by end of 2026, yet only **21% of organizations** maintain real-time inventory of their active agents [4].
- Organizations cannot wait for formal standards to appear. Effective governance of agentic systems requires immediate action on identity, authorization, audit logging, and human oversight structures — using frameworks that exist today.

Background

The arrival of commercially deployed agentic AI systems has created an accountability problem that standards bodies are only beginning to address. Unlike generative AI models that respond to discrete queries, AI agents are designed to act: they execute multi-step tasks, call external tools, read and write to data stores, spawn sub-agents, and operate continuously across enterprise environments with minimal

human interaction. The security implications are materially different in scope and operational consequence from those of passive AI systems, and the governance frameworks needed to address them are still being written.

NIST formalized that recognition on February 17, 2026, when its Center for AI Standards and Innovation announced the AI Agent Standards Initiative [1]. The initiative is organized around three strategic pillars: facilitating industry-led standards development through gap analyses and voluntary guidelines; fostering community-led open-source protocol development for agent interoperability with NSF investment; and investing in fundamental research on agent authentication, identity infrastructure, and security evaluation. The announcement was accompanied by a Request for Information (RFI) published in the Federal Register in January 2026 (docket NIST-2025-0035), seeking ecosystem perspectives on threat landscapes, existing mitigations, and measurement methodologies [6]. Comment periods on both the RFI and on the NCCoE's companion concept paper on AI agent identity and authorization have now closed.

That procedural progress, while meaningful, leaves a substantial near-term gap. The NIST Cybersecurity Framework Profile for Artificial Intelligence (NIST IR 8596) was released in preliminary draft form in December 2025, and the COSAiS project — which will produce SP 800-53 control overlays specifically for AI agents — published only a concept paper in August 2025 and an annotated outline for predictive AI use cases in January 2026 [2][7]. The overlays addressing single-agent and multi-agent deployments remain under active development with a projected publication timeline of late 2026 to 2027. Enterprises deploying agentic AI in 2026 are operating in advance of the controls specifically designed to govern them.

Security Analysis

A Standards Timeline That Trails Deployment Reality

The gap between deployment velocity and standards maturity is the defining feature of the current agentic AI security landscape. Enterprise adoption of AI agents has not slowed to wait for governance frameworks. Hyperscaler announcements, venture capital activity, and productivity mandates have created commercial pressure to deploy agentically at scale, while the technical bodies charged with producing the relevant controls — NIST, ISO/IEC JTC 1, and the NCCoE — are proceeding at the methodologically careful pace appropriate to standards work, a pace that creates a near-term gap relative to the speed of enterprise deployment.

The COSAiS project illustrates the asymmetry concretely. The concept paper released in August 2025 proposed five use cases for SP 800-53 overlay development, two of which address AI agent systems [2]. As of April 2026, the first overlay — for predictive AI, the least complex of the five use cases — exists only as an annotated outline. The overlays for both single-agent and multi-agent deployments are further back in the queue, with full publication not expected before late 2026 at the earliest and potentially into 2027. Organizations that deploy agentic systems today and wait for these overlays before building a control baseline will have operated without formal guidance for an extended period during which their agents are active, privileged, and exposed.

NIST's own empirical research makes the risk of that posture concrete. In January 2025, NIST researchers working in collaboration with the UK AI Security Institute conducted red-teaming exercises against an Anthropic Claude 3.5 Sonnet deployment using an enhanced version of the AgentDojo evaluation framework [3]. When red teamers applied novel attack strategies tailored to LLM-backed agent behavior — rather than baseline attack patterns known to the model — task-hijacking success rates rose from 11% to 81%. The attacks spanned three high-consequence categories: remote code execution via agent tool use, mass database exfiltration, and automated phishing conducted through agent communication channels. Crucially, successful attack techniques generalized across environments: methods developed against one configuration transferred to others even without detailed knowledge of the target setup, suggesting that the underlying vulnerabilities are architectural rather than configuration-specific.

The Identity and Credential Crisis

The NCCoE's February 2026 concept paper on "Accelerating the Adoption of Software and AI Agent Identity and Authorization" identified four core technical challenges: agent identification (distinguishing AI agents from human users), authorization (applying standards such as OAuth 2.0 to agent principals), access delegation (linking user identities to agents acting on their behalf), and logging and transparency (attributing agent actions to a non-human entity in a way that satisfies audit requirements) [5]. Each of these challenges has practical counterparts in the current enterprise environment that are not waiting for the concept paper to become a published project.

The CSA's 2026 Agentic Identity Survey — conducted in collaboration with Strata Identity and based on responses from 285 enterprise security practitioners — documents how organizations are actually managing agent credentials in the absence of standards [4]. Strata Identity is a commercial vendor in the agent identity management space; readers should evaluate the survey findings with awareness of that commercial context and the limited methodological disclosure available in the public summary. These findings suggest that the credential practices security teams have worked for a decade to eliminate from human authentication workflows are being replicated at scale in autonomous systems. Among

respondents, 44% authenticate agents using static API keys, 43% use username and password combinations, and 35% rely on shared service accounts — the same patterns that security operations teams have spent the past decade trying to remove from human authentication workflows, now reintroduced at scale into autonomous systems that operate continuously and with significant access privileges. Only 18% of security leaders expressed high confidence that their current IAM systems can effectively manage agent identities at all.

The visibility problem compounds the credential problem. Only 21% of organizations maintain a real-time inventory of their active agents, and only 28% can reliably trace agent actions back to a human sponsor across all environments [4]. When an agent takes an action — reads a file, calls an API, sends a message, spawns a sub-agent — the audit trail that would allow a security team to answer basic post-incident questions is absent in the majority of deployments. The NCCoE has identified this gap clearly. It has not yet resolved it.

Fragmented Accountability and the Ownership Question

The governance challenges documented in enterprise surveys are not only technical; they are organizational. The same survey data reveals that accountability for agent identity management is fragmented across security teams (39%), IT functions (32%), and specialized AI security functions (13%), with no dominant model of ownership [4]. This fragmentation means that no single organizational unit has clear responsibility for the agent security posture — a condition that commonly produces coverage gaps, as each function may assume other units are handling risks outside its immediate scope, a dynamic well-documented in organizational governance research.

The fragmentation also affects incident response. Security researchers have modeled scenarios in which a single compromised agent could influence downstream decision-making across connected agent processes within hours, though empirical baselines on propagation speed at enterprise scale remain limited. In an environment where most organizations cannot enumerate their active agents in real time, the ability to isolate a compromised agent before cascades propagate is severely constrained. Formal standards for agent behavior, state, and inter-agent communication — the kind that might enable automated circuit-breakers and isolation mechanisms — are among the deliverables that the NIST initiative aims to produce, but they are not yet available.

The industry feedback collected during NIST's comment periods offers some guidance on the trajectory of the standards themselves. Responding organizations broadly argued for flexible, voluntary frameworks over prescriptive mandates — an approach consistent with NIST's traditional model but one that places the burden of interpretation and implementation on organizations without definitive guidance [8]. A similar pattern played out in cloud security governance, where voluntary frameworks tended to

accelerate security improvements in organizations prepared to engage with them while leaving others without effective requirements until regulatory action created harder mandates — though the pace and scope of that transition varied significantly across sectors and jurisdictions.

Recommendations

Immediate Actions

Conduct an agent inventory before extending any new agent deployments. Organizations cannot govern what they cannot enumerate. Security teams should work with AI engineering, IT, and business units to produce a complete inventory of all currently active AI agents: their purpose, their identity credentials, the data and systems they can access, and the human sponsor responsible for each. This inventory is the prerequisite for every other governance action.

Replace static credentials with short-lived, scoped tokens for agent authentication. Static API keys and shared service accounts lack built-in expiration and are rarely rotated in practice, leaving the temporal window of a compromised credential effectively unconstrained without deliberate key management discipline. Organizations should migrate agent authentication to short-lived OAuth 2.0 tokens, scoped to the minimum necessary permissions for each agent's defined tasks, revocable on demand, and issued through an identity provider that maintains a machine-readable record of each grant.

Establish logging requirements for all agent actions at the tool-call level. Audit trails that capture only high-level outcomes are insufficient to support incident investigation or compliance reporting. Logging should capture each tool invocation, the data accessed or modified, the agent principal performing the action, and a timestamp — producing a record that can answer attribution questions after the fact even in multi-agent workflows.

Short-Term Mitigations

Apply MAESTRO threat modeling to all current agent deployments. CSA's MAESTRO framework provides a seven-layer reference architecture specifically designed for agentic AI threat analysis, addressing vulnerabilities across Foundation Models, Data Operations, Agent Frameworks, Deployment and Infrastructure, Evaluation and Observability, and the Agent Ecosystem [9]. Applying MAESTRO to

existing deployments will surface risks — particularly agentic-specific threats like goal misalignment, prompt injection via tool results, and trust boundary violations — that conventional threat modeling frameworks such as STRIDE were not designed with agentic architectures in mind to capture.

Implement human-in-the-loop checkpoints for high-consequence actions. Organizations should define categories of high-consequence agent action — access to production data stores, financial transactions above a threshold, external communications, sub-agent spawning — and require human approval before agents execute in those categories. This applies the organizational equivalent of a compensating control for the missing technical guardrails while formal standards are developed.

Engage with the NIST COSAiS development process. The COSAiS Slack community and the `overlays-securing-ai@list.nist.gov` contact channel remain active venues for practitioner input into the overlay development process; readers should verify current engagement channels through the CSRC project page [2]. Organizations with mature agent deployments are particularly well-positioned to contribute practical guidance on the single-agent and multi-agent use cases, and their engagement will improve the quality of the eventual controls.

Strategic Considerations

Build toward identity infrastructure that scales to 2027 and beyond. The projected growth of non-human and agentic identities — toward tens of billions by the end of 2026 — makes identity governance the long-term constraint on secure agent deployment [4]. Organizations should begin evaluating identity orchestration platforms and agent-aware IAM tooling now, treating agent identity as a first-class security problem rather than an extension of existing service account management.

Anticipate regulatory codification of agent governance requirements. The trajectory of past voluntary federal cybersecurity frameworks — including the NIST Cybersecurity Framework's progressive incorporation into sector-specific regulatory requirements — suggests that mandatory requirements often follow voluntary guidance, particularly following high-profile incidents. Organizations that build agent governance programs now — against MAESTRO, the forthcoming COSAiS overlays, and the NCCoE's identity and authorization guidance — will be positioned to demonstrate compliance posture when regulatory requirements crystallize. Organizations that defer governance investment until mandates arrive are likely to face compressed implementation timelines with less operational flexibility to phase in controls.

Develop inter-team accountability structures before agents require them. The fragmented ownership observed in current enterprise deployments is a governance risk independent of any specific technical vulnerability. Security, IT, AI engineering, legal, and business functions should establish

documented responsibilities for the lifecycle of AI agents: who approves deployment, who holds credentials, who monitors behavior, who responds to incidents, and who owns remediation. These accountability structures should be established before incidents demonstrate their necessity.

CSA Resource Alignment

As the authoring organization, CSA highlights its own frameworks here as applicable resources; practitioners should also evaluate complementary frameworks including MITRE ATLAS and ISO/IEC 42001 when building comprehensive governance programs.

CSA's published frameworks provide enterprises with actionable governance foundations that do not require waiting for the NIST standards pipeline to complete. The **MAESTRO framework** (Multi-Agent Environment, Security, Threat, Risk and Outcome) offers the most direct alignment, providing a structured threat modeling methodology tuned to the layered architecture of agentic AI systems and explicitly addressing the novel threat classes — non-determinism, autonomy, and trust boundary violations — that conventional frameworks leave unaddressed [9]. Security teams working to assess existing agent deployments should treat MAESTRO as a starting point.

CSA's **AI Organizational Responsibilities** guidance addresses the governance and accountability dimension that is currently underspecified in the NIST standards process, providing a framework for assigning and documenting human responsibility for AI systems across organizational functions. This guidance is directly applicable to the fragmented ownership problem documented in current enterprise surveys. The **AI Controls Matrix (AICM)** — a superset of the Cloud Controls Matrix (CCM) that extends its control mappings with AI-specific requirements — provides coverage across identity and access management, logging, and incident response that can be applied to agentic contexts while organizations await the formal COSAiS overlays. CSA's **STAR program** offers a mechanism for organizations to demonstrate their agent governance posture to third parties, building toward the audit and assurance infrastructure that will be required when regulatory requirements follow the standards work currently underway at NIST.

References

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

[2] NIST Computer Security Resource Center. "SP 800-53 Control Overlays for Securing AI Systems (COSAiS)." CSRC, accessed April 2026.

[3] NIST CAISI. "Technical Blog: Strengthening AI Agent Hijacking Evaluations." NIST, January 17, 2025 (updated December 2025).

[4] CSA and Strata Identity. "The AI Agent Identity Crisis: New Research Reveals a Governance Gap." Strata Identity, 2026.

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

[6] Federal Register / NIST. "AI Agent Security Request for Information (NIST-2025-0035)." Federal Register, January 2026.

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

[8] Cybersecurity Dive. "Industry to NIST: Keep Agentic AI Standards Flexible and Voluntary." Cybersecurity Dive, 2026.

[9] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA, February 2025.

