CSAI Foundation

Cloud Security Alliance AI Safety Initiative

The AI Agent Governance Gap: What CISOs Need Now

NIST's Standards Initiative, the CAISI RFI, and Immediate CISO Priorities for Agentic AI

Unofficial AI-assisted Research

2026-04-03

Key Takeaways

- NIST's Center for AI Standards and Innovation (CAISI) issued a Request for Information on January 8, 2026 — the first formal U.S. government initiative specifically scoped to cybersecurity controls for autonomous AI agent systems [1].
- The NIST AI Agent Standards Initiative, announced February 17, 2026, represents a multi-year standards effort; no enforceable, agent-specific security controls exist today, and the first substantive NIST deliverables are not expected before late 2026 at the earliest [2].
- According to the CSA State of AI Cybersecurity 2026 survey of over 1,500 security leaders, 92% of organizations are concerned about AI agent security implications, yet most organizations report significant gaps in comprehensive AI security governance [3].
- Existing frameworks — including NIST AI RMF 1.0, ISO/IEC 42001:2023, and the EU AI Act — were architected before the era of autonomous, tool-calling agents and contain structural gaps that complicate direct application to agentic deployments [4][5][6].
- The identity and authorization gap is operationally acute: a 2026 survey of 235 large-enterprise CISOs and CIOs found that 92% lack full visibility into their AI agent identities, and 95% doubt they could detect or contain a compromised agent [7].
- CISOs cannot wait for standards bodies. Organizations must establish internal governance for agentic AI now, drawing on available near-term resources including the OWASP Agentic Top 10, the NCCoE AI agent identity concept paper, and CSA's AI Controls Matrix (AICM).

Background

The Rapid Arrival of Agentic AI in the Enterprise

AI agent systems are categorically different from the AI tools that informed the governance frameworks organizations use today. Unlike earlier AI assistants that responded to a prompt and returned a single output, an AI agent autonomously plans, selects tools, executes multi-step tasks, and adapts its behavior in response to environmental feedback — often without human intervention at each step. Multi-agent

architectures extend this further: a coordinating "orchestrator" agent delegates subtasks to specialized subordinate agents, each with access to external APIs, data stores, code execution environments, and communication systems. The result is an AI capability that acts, not merely responds.

Enterprise adoption of this capability is accelerating at a pace that governance has not matched. Gartner projects that 40% of enterprise applications will embed task-specific AI agents by end of 2026, up from fewer than 5% in 2025 [8]. This trajectory means that organizations across financial services, healthcare, critical infrastructure, and technology are deploying systems that can autonomously query databases, send emails, execute code, and modify cloud configurations — often with the same permissions as the human employees who provisioned them. Security leaders have taken notice: the Cisco/Splunk CISO Report, based on 650 global CISOs surveyed in mid-2025, found that 86% fear agentic AI will increase social engineering attack surface, and 82% worry about faster adversarial persistence mechanisms enabled by AI autonomy [9].

The Standards Landscape Before 2026

Prior to this year, organizations seeking governance guidance for agentic AI were directed toward frameworks designed for a fundamentally different technology paradigm. NIST AI RMF 1.0, published in January 2023, provides a risk management vocabulary and governance process for AI development and deployment. The AI RMF 1.0 was designed for AI systems whose behavior can, in principle, be characterized at deployment time and whose decisions can be subject to human review — conditions that autonomous, tool-calling agents routinely violate by operating at machine speed across dynamic environments [4]. ISO/IEC 42001:2023, the world's first certifiable AI management system standard, provides a plan-do-check-act governance structure with 38 controls, but was similarly designed as a general framework for AI system management rather than real-time policy enforcement for agentic architectures [5]. NIST IR 8596, published in December 2025 as an initial public draft, offers a cybersecurity framework profile for AI systems and represents meaningful progress; however, it remains in draft form and does not address the specific control requirements for autonomous agents executing multi-step tasks across distributed tool ecosystems [19]. The EU AI Act, which began enforcing general-purpose AI obligations in August 2025, contains no definition of "agentic systems." One analysis argues that key provisions — including Article 43 (Conformity Assessment), Article 9 (Risk Management), and Article 14 (Human Oversight) — were drafted on the premise that AI system behavior is known, documentable, and stable at deployment time, an assumption autonomous agents invalidate by design [6][18].

The gap was not hypothetical. By early 2026, security researchers had documented approximately 8,000 MCP servers exposed on the public internet without authentication, creating direct attack surface through the very tool-calling infrastructure that agentic systems depend upon [10]. Separately, security

researchers documented over 30 vulnerabilities in the Model Context Protocol ecosystem within a 60-day window, a rate of discovery that underscored how quickly the agentic attack surface was expanding [10]. Prompt injection — ranked first on the OWASP Top 10 for Large Language Model Applications since 2025 — represents the most operationally significant attack surface in production LLM systems [11]. The standards community had identified the problem; what was absent was actionable, agent-specific guidance.

Security Analysis

The CAISI RFI: Official Acknowledgment of a Structural Gap

On January 8, 2026, CAISI [20] published docket NIST-2025-0035 in the Federal Register: *"Request for Information Regarding Security Considerations for Artificial Intelligence Agents."* The RFI's scope explicitly acknowledged what practitioners had documented for months — that conventional cybersecurity approaches do not translate cleanly to autonomous agent deployments. CAISI solicited input across seven domains: gaps in existing cybersecurity frameworks when applied to AI agents; threat models and vulnerabilities specific to autonomous task execution; methods for measuring agent system security during development; deployment interventions to constrain and monitor agent access; governance and oversight controls for production environments; secure development lifecycle practices adapted for agentic systems; and approaches to monitoring, logging, and incident response [1].

The comment period closed March 9, 2026. The public record of docket NIST-2025-0035 reflects responses from academic research centers including the UC Berkeley Center for Long-Term Cybersecurity, industry bodies including the Business Software Alliance, and numerous enterprise security teams [1]. The substantive breadth of that response confirmed that practitioners were encountering real governance failures, not theoretical edge cases. The RFI itself does not produce standards; it is the input-gathering phase preceding a multi-year standards development effort that NIST formally inaugurated six weeks later.

NIST's Three-Pillar Response and Its Timeline

The AI Agent Standards Initiative, announced February 17, 2026, organizes NIST's response around three strategic pillars: industry-led standards facilitation through technical convenings and gap analyses; community-driven development of open-source interoperability protocols — with the Model Context Protocol (MCP) and the emerging Agent-to-Agent (A2A) protocol identified as interoperability baselines — targeting an AI Agent Interoperability Profile by Q4 2026; and fundamental research on

agent authentication, identity infrastructure, and security evaluations [2]. SP 800-53 control overlays specifically designed for single-agent and multi-agent AI systems are described as forthcoming but remain in development as of this writing.

The initiative's timeline creates a meaningful planning horizon for security practitioners: the first substantive NIST deliverables are a year or more away, agent-specific SP 800-53 overlays are further still, and the international standards process at ISO/IEC JTC 1 operates on timelines measured in years. Organizations deploying AI agents today are doing so in the absence of established standards, which makes the CAISI RFI's publication strategically important as a signaling event even before it produces guidance. It represents the U.S. government's formal acknowledgment that the governance gap is real, that it requires dedicated standards work, and that existing frameworks are insufficient substitutes.

The Identity and Authorization Gap

One of the most immediately exploitable dimensions of the agentic governance shortfall is the failure of enterprise identity infrastructure to account for non-human principals operating at scale. AI agents present a novel identity challenge: they act on behalf of users, hold delegated credentials, make authorization decisions in real time, and may spawn sub-agents with their own permission sets — all with no coherent, purpose-built mechanism in existing IAM frameworks for representing the delegated, multi-level, and real-time authorization patterns that agentic systems require.

The NCCoE's concept paper published February 5, 2026, *"Accelerating the Adoption of Software and AI Agent Identity and Authorization,"* describes a future project that would apply existing identity standards — OAuth 2.0 extensions, SP 800-207 Zero Trust Architecture, SP 800-63-4 Digital Identity Guidelines — to AI agent scenarios [12]. That this is a concept paper, not published guidance, illustrates the state of the field: even the identity-specific problem has not yet been addressed by a completed framework. The 2026 CISO AI Risk Report captures the operational consequence: among 235 large-enterprise security leaders, 92% lack full visibility into their AI identities, 86% do not enforce access policies for AI identities, and 71% report that AI systems have access to core business platforms — ERP, CRM, and financial systems — while only 16% govern that access effectively [7]. Industry research consistently shows that non-human identities — including AI agents, service accounts, and automation bots — already outnumber human identities in most large enterprise environments, yet most organizations' PAM and IAM tooling was not architected with autonomous, task-executing AI principals in mind.

The Audit Opacity Problem

Autonomous agents acting across distributed tool ecosystems generate a fundamentally different audit trail than human users or traditional automated workflows. An agent might call a dozen tools, spawn multiple sub-agents, read from a vector database, and write to a production API — all within a single task execution — with the intermediate reasoning states that determined those actions remaining inside the model and inaccessible to conventional logging. Organizations that have invested in SIEM pipelines, DLP controls, and user behavioral analytics find those controls provide only partial coverage: they may capture the terminal API call but miss the chain of agent decisions that produced it.

The data reflects a consequential monitoring gap. According to the EY/AIUC-1 Consortium survey published in Help Net Security in March 2026, only 38% of organizations monitor AI traffic end-to-end across prompts, tool calls, and outputs; only 17% continuously monitor agent-to-agent interactions [13]. That same survey found that 64% of companies with revenue above $1 billion reported losses exceeding $1 million that they associated with AI system failures during 2025, and that 80% of surveyed organizations documented risky agent behaviors including unauthorized system access and data exposure. The absence of evidence-quality audit trails is both a security problem and a compliance liability: regulators examining AI-involved incidents will expect organizations to reconstruct what their agents did, why, and with whose authorization.

What the Regulatory Vacuum Means in Practice

The absence of agent-specific regulatory requirements is sometimes misread as permission to defer governance work. The more accurate reading is that organizations face an especially difficult governance condition: unclear rules today, with enforcement pressure building from multiple directions. The EU AI Act's August 2026 enforcement deadline for high-risk AI systems will arrive before agent-specific guidance does; EU regulators applying existing provisions to agentic deployments will make interpretive decisions that may or may not favor the organization under review. In the United States, the White House's National AI Legislative Framework, released March 20, 2026, adopts a light-touch, sector-based approach and does not establish autonomous AI agents as a distinct regulatory category [14]. However, sector regulators — the OCC, FFIEC, FDA, CISA, and SEC among others — have existing authorities they may apply to AI agent deployments within their jurisdictions, and enforcement discretion is not the same as regulatory absence.

The practical implication is that organizations in regulated industries should conduct AI agent inventories and governance gap assessments now, without waiting for explicit AI-agent-specific rules to appear. The governance infrastructure built today — agent registries, authorization policies, audit log pipelines,

incident response playbooks — will be the artifact an organization presents to regulators when agent-specific enforcement arrives.

Recommendations

Immediate Actions

Establish an AI agent inventory. Organizations cannot govern what they cannot see. Security teams should enumerate all AI agent deployments, including those provisioned by business units without formal IT involvement. The inventory should capture agent identity, delegated permissions, connected tools and data sources, human owner of record, and the business process each agent supports. This is the prerequisite for every subsequent governance measure and is an asset an organization will need to produce in any future regulatory examination.

Apply least-privilege to agent credentials immediately. Even in the absence of agent-specific identity standards, existing PAM and IAM controls can be applied. Agents should not hold standing access to production systems; where possible, they should receive time-bound, just-in-time credentials scoped to the specific resources their task requires. The 2026 CISO AI Risk Report finding that only 16% of organizations effectively govern AI access to core business systems indicates that most organizations have significant room for improvement using existing tooling [7].

Begin monitoring agent-to-agent traffic. Organizations should extend their existing SIEM and UEBA architectures to capture AI agent interactions, including tool call sequences, external API requests, and data access patterns. Even imperfect coverage is substantially better than none: anomalous agent behavior detected after the fact is recoverable; behavior that leaves no log is not.

Short-Term Mitigations

Adopt the OWASP Agentic Top 10 as a baseline threat model. Published December 10, 2025, the OWASP Agentic Top 10 is the most operationally actionable agent-specific security framework currently available, covering the primary attack surfaces unique to autonomous agent systems — including agent goal hijacking, agentic supply chain vulnerabilities, unexpected code execution, memory and context poisoning, insecure inter-agent communication, and rogue agent behavior [15]. Security teams should map their agent deployments against this taxonomy to identify unaddressed risks and prioritize controls work.

Apply the NCCoE AI Agent Identity concept paper as an architectural guide. While the NCCoE's February 2026 concept paper is not yet published guidance, its proposed application of OAuth 2.0, Zero Trust (SP 800-207), and Digital Identity Guidelines (SP 800-63-4) to agent scenarios provides a viable architectural blueprint for organizations that need to make identity infrastructure decisions before standards are finalized [12]. Teams implementing agent authorization frameworks should document their design decisions and the standards they reference, creating a compliance-ready audit trail.

Implement agent-specific incident response procedures. Standard IR playbooks assume human actors or traditional malware; neither translates well to a compromised AI agent that may be autonomously executing harmful actions across dozens of connected systems. Organizations should develop agent-specific IR procedures that address how to revoke agent credentials, how to isolate an agent from its tool ecosystem, how to reconstruct the sequence of agent actions during an incident, and what notification obligations may apply.

Strategic Considerations

Monitor the CAISI RFI response outcomes and AI Agent Standards Initiative deliverables. NIST has committed to publishing an AI Agent Interoperability Profile by Q4 2026 and is developing SP 800-53 control overlays for agentic systems. Organizations should treat these deliverables as inputs to their governance frameworks and plan to adopt them when published, both because they represent best available guidance and because regulatory expectations will increasingly reference them.

Engage in the standards process. IANS Research found in February 2026 that approximately 50% of large enterprises have established dedicated AI governance committees [16]. Organizations with technical security expertise should use these committees to engage with NIST, ISO/IEC JTC 1, and industry bodies developing agent-specific standards, both to influence the outcome and to develop internal expertise that will be essential for implementation.

Prepare for agent-specific compliance examinations. Organizations in regulated industries should plan for existing sector regulations to be interpreted to cover AI agent deployments within the next 12 to 18 months, based on current regulatory trajectories across both the EU and U.S. sector agencies. Building governance documentation, agent inventories, and audit log infrastructure now positions the organization for compliance rather than remediation under pressure.

CSA Resource Alignment

CSA's AI Safety Initiative has published several frameworks directly applicable to the governance challenges described in this note. The **AI Controls Matrix (AICM)** [21], CSA's primary governance framework for AI systems, provides 18 security domains and over 240 control objectives mapped to the full AI lifecycle, covering roles including AI customers, orchestrated service providers, model providers, and application providers — a taxonomy that directly addresses the multi-party accountability questions that agentic deployments raise. The **Capabilities-Based Risk Assessment (CBRA) for AI Systems** methodology applies a multiplicative scoring framework across four dimensions — System Criticality, AI Autonomy, Access Permissions, and Impact Radius — that provides a principled approach to prioritizing governance investment across an organization's agent portfolio.

CSA's **MAESTRO** threat modeling framework provides structured threat analysis for multi-agent AI architectures, covering the specific attack surfaces — orchestrator compromise, sub-agent hijacking, tool ecosystem poisoning — that make agentic systems categorically different from static AI deployments. Organizations conducting agent security assessments should apply MAESTRO in conjunction with the OWASP Agentic Top 10 to ensure comprehensive threat coverage. The **CSA STAR** program's assurance mechanisms, and the extended assurance capabilities planned under the new **CSAI Foundation** announced March 23, 2026, will provide organizations a structured pathway for demonstrating agent governance to customers, partners, and regulators as this space matures [17].

The **AI Organizational Responsibilities** series — covering governance, risk management, compliance, and core security responsibilities — provides the organizational and cultural governance layer that technical controls alone cannot substitute. For organizations standing up AI agent governance programs, these publications provide the board-level framing and role-assignment structure that technical teams require to operate with clear mandates and accountability.

References

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

[2] NIST. "Announcing the AI Agent Standards Initiative: Interoperable and Secure AI Agents". NIST News. February 17, 2026.

[3] Cloud Security Alliance. "The State of AI Cybersecurity 2026: Unveiling Insights from Over 1,500 Security Leaders". CSA Blog. April 2, 2026.

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

[5] ISO/IEC. "ISO/IEC 42001:2023 — Artificial Intelligence Management Systems". International Organization for Standardization. December 2023.

[6] Hannecke, M. "Agentic Tool Sovereignty: The EU AI Act Problem Nobody Saw Coming". Medium. January 27, 2026.

[7] Cybersecurity Insiders. "2026 CISO AI Risk Report". Cybersecurity Insiders. January 24, 2026.

[8] Gartner. "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up From Less Than 5% in 2025". Gartner Newsroom. August 26, 2025.

[9] Cisco/Splunk. "Splunk Report: Agentic AI Takes Center Stage in CISOs' Path to Digital Resilience". Cisco Newsroom. February 2026.

[10] Cikce. "8,000 MCP Servers Exposed: The Agentic AI Security Crisis of 2026". Medium. 2026.

[11] OWASP. "OWASP Top 10 for Large Language Model Applications — 2025". OWASP. 2025.

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

[13] Help Net Security. "Enterprise AI Agent Security 2026". Help Net Security / EY–AIUC-1 Consortium Survey. March 3, 2026.

[14] Nixon Peabody. "White House Releases National AI Legislative Framework". Nixon Peabody Insights (secondary analysis). March 26, 2026.

[15] OWASP GenAI. "OWASP Top 10 for Agentic Applications 2026". OWASP GenAI Security Project. December 10, 2025.

[16] IANS Research. "The CISO's Expanding AI Mandate: Leading Governance in 2026". IANS Research Blog. February 6, 2026.

[17] Cloud Security Alliance. "CSA Launches AI Safety Initiative Foundation at RSA Conference 2026". CSA Press Release. March 23, 2026.

[18] Berkeley Center for Long-Term Cybersecurity. "Agentic AI Risk Management Standards Profile". UC Berkeley CLTC. February 2026.

[19] NIST. "NIST IR 8596 (Initial Public Draft): Cybersecurity Profile for Artificial Intelligence (Cyber AI Profile)". NIST CSRC. December 2025.

[20] NIST CAISI. "Center for AI Standards and Innovation". NIST. 2026.

[21] Cloud Security Alliance. "AI Controls Matrix (AICM)". CSA AI Working Group. 2025–2026.