# Governing the Agent: NIST's AI Agent Standards Initiative

Emerging Regulatory Frameworks for Autonomous AI Systems

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-23

 

## Key Takeaways

- NIST's Center for AI Standards and Innovation (CAISI) launched the AI Agent Standards Initiative on February 17, 2026, establishing three pillars: industry-led standards development, community-led open-source protocol development, and foundational research on AI agent security and identity [1].
- The initiative follows a January 2026 Request for Information that received significant public comment and explicitly scoped its focus to agents capable of creating "persistent changes outside of the AI agent system itself" — distinguishing agentic systems from general-purpose AI chatbots and retrieval-augmented generation [2][3].
- NISTIR 8596, a preliminary draft released in December 2025, establishes a Cybersecurity Framework Profile for AI that addresses agentic-specific risks including excessive autonomy, model drift, and human oversight deficits [4].
- The EU AI Act contains no provisions specifically addressing multi-step autonomous decision-making or agent orchestration; the European Parliament's Committee on Internal Market and Consumer Protection (IMCO) and Committee on Civil Liberties, Justice and Home Affairs (LIBE) jointly backed a proposal on March 18–19, 2026 to delay full high-risk AI obligations by 16 months, to December 2, 2027 [5][6].
- Singapore's Infocomm Media Development Authority (IMDA) published what it described as the world's first governance framework specifically designed for AI agents in January 2026, establishing four governance dimensions and addressing risk categories including erroneous actions, unauthorized actions, biased decisions, data breaches, and disruption to connected systems [7].
- Security practitioners should not wait for regulatory finalization: CSA's AI Controls Matrix (AICM) [10], Agentic AI Red Teaming Guide [11], and AI Organizational Responsibilities series [12] provide actionable, framework-aligned controls for managing agentic AI risk today.

# Background

The widespread deployment of AI agents — systems that perceive their environment, plan multi-step actions, and execute tasks with varying degrees of human oversight — has exposed a significant gap in the global regulatory landscape. As of early 2026, no major standards body had published guidance designed specifically for the agentic use case — a gap CAISI's AI Agent Standards Initiative was explicitly organized to address [1][2]. Traditional cybersecurity frameworks addressed software systems and data pipelines; AI governance frameworks addressed model development and deployment; but the intersection of autonomous decision-making, tool use, and real-world action remained largely uncharted territory for regulators and standards developers alike.

This gap reflects the pace of capability development outrunning the slower cycles of standards development rather than any failure of attention. Agentic AI systems occupy a novel category that challenges several assumptions underlying existing frameworks. Unlike static AI models evaluated at a point in time, agents evolve during operation, accumulate context, invoke external tools and APIs, and may spawn sub-agents that inherit or escalate permissions. Their behavior emerges from the interaction of a model, its context window, and the environment in which it operates — making pre-deployment testing insufficient as a standalone assurance mechanism. The combination of autonomy and real-world consequence distinguishes these systems categorically from the chat interfaces and image classifiers that dominated earlier AI governance discussions.

NIST's Center for AI Standards and Innovation (CAISI) has emerged as NIST's primary vehicle for coordinating U.S. government engagement on AI agent standards, with a mandate that includes developing voluntary guidelines for AI systems, conducting unclassified evaluations focused on demonstrable risks including cybersecurity and biosecurity, and representing U.S. interests in international AI standards bodies. The January 2026 Request for Information and the subsequent AI Agent Standards Initiative reflect CAISI's recognition that the agentic AI category requires dedicated, coordinated attention distinct from the AI governance work that preceded it.

The international dimension of this challenge is equally significant. The EU AI Act, which entered into force in August 2024 and has been applying obligations in stages since February 2025, does not define "agentic AI systems." Singapore's IMDA was among the first jurisdictions to publish framework-specific guidance for agentic AI, moving ahead of most other regulatory bodies. The resulting regulatory environment is fragmented, with different jurisdictions at different stages of framework maturity and some of the most consequential obligations still pending finalization or subject to proposed delays.

---

# Security Analysis

## The NIST AI Agent Standards Initiative

CAISI's February 17, 2026 announcement described the AI Agent Standards Initiative as an effort to ensure that AI agents "capable of autonomous actions" can be adopted with confidence, function securely on behalf of users, and interoperate across the digital ecosystem [1]. The initiative is organized around three pillars that reflect the distinct dimensions of the agentic AI governance problem.

The first pillar, industry-led standards development, tasks CAISI with facilitating technical convenings, conducting gap analyses of existing standards, producing voluntary guidelines, and strengthening U.S. leadership in international standards bodies. This work reflects community feedback that the current standards landscape — including NIST AI RMF, NIST SP 800-53, and the emerging ISO/IEC 42001 framework for AI management systems — does not adequately address the specific risk surface created by agents that take autonomous, real-world actions. A National Science Foundation investment in open-source ecosystems through the Pathways to Enable Secure Open-Source Ecosystems program supports the second pillar, which focuses on identifying and removing barriers to interoperable agent protocols.

The third pillar's focus on AI agent security and identity research addresses what may be the least mature area of current practice: how agents authenticate, how permissions are scoped, and how actions are audited. CAISI is advancing fundamental research on agent authentication and identity infrastructure — a problem without adequate solutions in most current deployments. In the absence of dedicated identity standards for agents, deployments often lack clearly defined identity chains, robust authentication mechanisms, and comprehensive audit trails for consequential agent actions. When an agent acts on behalf of a user, makes API calls, or spawns subordinate agents, the identity chain is poorly understood and poorly auditable in practice — a gap that the NCCoE's concurrent Draft Concept Paper on "Accelerating the Adoption of Software and AI Agent Identity and Authorization" is scheduled to address when it opens for public comment on April 2, 2026 [1].

## The January 2026 RFI: Defining the Threat Surface

The CAISI RFI published in the Federal Register on January 8, 2026 is notable as much for what it excluded as for what it addressed [3]. By explicitly limiting its scope to agents capable of creating "persistent changes outside of the AI agent system itself" — and by excluding general generative AI, chatbots, and retrieval-augmented generation from its scope — CAISI made a deliberate definitional choice that will likely shape U.S. regulatory frameworks for years. The implication is clear: an AI system

that merely retrieves information or generates text is categorically different, in a governance and security sense, from one that can schedule meetings, execute code, transfer funds, modify database records, or interact with industrial control systems.

The significant public comment received in response to the RFI [2] reflects broad stakeholder engagement across industry, government, and civil society with this question. CAISI's framing of the core threat categories is instructive. Indirect prompt injection attacks — in which malicious content in the agent's environment manipulates its behavior without the user's knowledge — represent an adversarial risk with no clean analogue in traditional software security. Data poisoning attacks that corrupt training data, specification gaming in which agents pursue misaligned objectives, and hybrid vulnerabilities arising specifically from combining model outputs with software system functionality are all threat categories that existing security frameworks address only partially or not at all.

CAISI also asked respondents to address how threats evolve as agent capabilities expand — acknowledging that the security posture required today may prove insufficient for more capable agents emerging over the near term, a consideration relevant to enterprise risk planning timelines. This forward-looking framing is consistent with the pace of capability development in this area.

## NISTIR 8596: Cybersecurity Framework Profile for AI

The December 2025 preliminary draft of NISTIR 8596, the Cybersecurity Framework Profile for Artificial Intelligence, provides the most detailed NIST guidance to date on managing AI-specific cybersecurity risks within the familiar CSF 2.0 structure [4]. The Cyber AI Profile organizes guidance across three missions: securing AI systems from cyber threats, conducting AI-enabled cyber defense, and thwarting AI-enabled cyberattacks. Each mission is mapped to the CSF 2.0 functions — Govern, Identify, Protect, Detect, Respond, and Recover — creating a navigable structure for organizations already operating within the CSF.

For agentic systems specifically, the draft establishes guidance that security teams should treat as current best practice regardless of the finalization timeline. NIST describes AI systems as "autonomous entities capable of interacting with data, systems, and even other agents, sometimes at machine speed and without direct human intervention," and directs organizations to assign clear human accountability for every such system [4]. The draft explicitly identifies excessive autonomy as a risk: if a task does not genuinely require autonomous action, architecting it that way expands the attack surface without adding value. Human-in-the-loop checkpoints, model drift monitoring, and maturity assessments before deployment are identified as core controls for agentic systems.

## The EU Regulatory Picture: Provisions, Gaps, and Delays

The EU AI Act presents a structurally complex situation for organizations deploying AI agents. The Act does not define agentic AI systems, and as of this writing the EU AI Office has not published dedicated guidance addressing agentic deployments specifically, indicating that regulatory considerations in this area remain at an early stage [6]. However, agentic deployments are not unregulated under the Act; they are subject to two overlapping pathways that apply based on deployment context and model characteristics.

Agents deployed in high-risk domains enumerated in Annex III — including critical infrastructure, employment and worker management, essential services, law enforcement, and administration of justice — trigger the full suite of high-risk AI obligations. These include lifecycle-spanning risk management systems, data governance requirements, technical documentation, transparency and human oversight mechanisms, and post-market monitoring. For agents powered by general-purpose AI models whose training compute exceeds $10^{25}$ FLOPs, the systemic risk provisions applicable since August 2, 2025 require adversarial testing, systemic risk assessment and mitigation, and incident reporting to the EU AI Office.

The enforcement timeline for the most substantive obligations is in flux. On March 18–19, 2026, the European Parliament's Committee on Internal Market and Consumer Protection (IMCO) and Committee on Civil Liberties, Justice and Home Affairs (LIBE) jointly adopted Report A10-0073/2026 by a vote of 101 in favor, 9 against, and 8 abstentions [5]. The proposal would move the Annex III high-risk system obligations from August 2, 2026 to December 2, 2027, and push Annex I high-risk obligations to August 2, 2028. A six-month extension on generative AI watermarking requirements under Article 50(2) is also included. These are committee-level proposals pending full Parliament and Council processes, but they signal significant political momentum toward delay. Organizations should plan compliance programs against the revised timeline while monitoring formal adoption.

## The Singapore Model: Purpose-Built Agentic Governance

Singapore's IMDA published what it characterized as the world's first governance framework specifically designed for AI agents at the World Economic Forum in Davos in January 2026 [7]. While voluntary and non-binding, the framework's architecture is instructive precisely because it was designed with the agentic use case in mind rather than adapted from existing AI governance constructs.

The framework is organized around four governance dimensions. Risk assessment and bounding requires use-case-specific analysis that accounts for the agent's autonomy level, data sensitivity, and action scope — recognizing that a scheduling agent poses categorically different risks than an agent authorized to act on financial systems. Human accountability obligations require clear organizational responsibility

allocation across developers, deployers, operators, and end users, with mandatory human override and intercept checkpoints. Technical controls span the full lifecycle from design guardrails through pre-deployment testing to progressive rollout and real-time monitoring. End-user responsibility provisions address transparency about agent capabilities and escalation pathways.

The risk categories the framework addresses — including erroneous or unauthorized actions, biased or unfair decisions, data breaches and inadvertent disclosure, and disruption to connected systems — map closely to the threat categories identified in CAISI's RFI and CSA's own Agentic AI Red Teaming Guide [7]. The alignment between these independently developed frameworks — one from a U.S. federal agency, one from Singapore's digital authority — suggests these threat categories reflect broadly shared practitioner understanding of the agentic AI risk surface, though the frameworks share common source literature.

### CISA Context: Reduced Capacity, Persistent Guidance

CISA's website noted a lapse in federal funding at the time of this writing, with an advisory that the site would not be actively managed during that period [8]. Despite this operational constraint, CISA's published AI guidance remains in force and directly relevant to organizations deploying agents in or adjacent to critical infrastructure. The December 2025 joint guidance with NSA, FBI, and international partner agencies on principles for secure integration of AI in operational technology, the AI Data Security best practices document, and the AI Cybersecurity Collaboration Playbook issued through the Joint Cyber Defense Collaborative all provide baseline security expectations that inform responsible agentic AI deployment [9].

---

## Recommendations

### Immediate Actions

Organizations currently deploying or evaluating AI agents should treat the NIST, EU, and Singapore frameworks not as a reason to wait but as a basis for action now. The CAISI RFI's threat taxonomy — indirect prompt injection, data poisoning, specification gaming, and hybrid vulnerabilities from combining AI outputs with software functionality — defines the threat categories that security teams need to assess today. CSA's Agentic AI Red Teaming Guide, which addresses twelve distinct threat categories for autonomous agents including authorization hijacking, checker-out-of-the-loop vulnerabilities, and multi-agent exploitation, provides a structured methodology for conducting that assessment before deployment.

The identity and authorization gap is the most pressing immediate concern. Organizations should audit current agent deployments to determine how agents authenticate, what permissions they hold, how those permissions are scoped and revoked, and whether an end-to-end audit trail is maintained for consequential agent actions. If the answer to any of these questions is unclear, that represents a material security gap regardless of where the regulatory frameworks eventually settle.

## Short-Term Mitigations

The NISTIR 8596 draft's guidance on agentic system design principles should be incorporated into AI procurement and development standards now. The principle that agentic architecture should be deployed only where genuine autonomy is required, and that every agent deployment requires identified human accountability, provides clear criteria for evaluating proposed deployments. Organizations should implement human-in-the-loop checkpoints proportionate to the consequence level of agent actions, with more frequent intervention points for agents operating in high-stakes or irreversible decision contexts.

For organizations subject to EU AI Act obligations, the delay proposals under consideration should be treated as planning information rather than a guarantee of relief. Compliance programs structured around the original August 2026 timeline remain the prudent approach given that committee-level approval does not guarantee final adoption, and organizations that build compliance capability earlier will face less disruption regardless of the final enforcement date. The trajectory of EU AI Office activity suggests that agentic-specific provisions may emerge before the broader implementation delays resolve, as the Act's existing risk-classification structure provides a basis for extending obligations to agentic deployments without legislative amendment.

## Strategic Considerations

The convergence of the NIST AI Agent Standards Initiative, the EU AI Act agentic guidance gap, and Singapore's first-mover governance framework reflects a global recognition that agentic AI systems require dedicated governance attention. Organizations with significant AI agent deployments should engage with the standards development process directly – through CAISI's listening sessions on sector-specific adoption barriers, through comment processes on the NCCoE concept paper, and through industry association channels in EU standards bodies. Organizations that shape these frameworks will be better positioned to meet the resulting obligations than those that engage only when final rules are published.

The trajectory of the international regulatory landscape points toward mandatory requirements for agent identity, authorization scoping, human accountability assignment, and incident reporting for consequential agent failures. Building those capabilities now, as recommended by existing voluntary frameworks, creates both security value and a head start on compliance obligations that appear likely to become mandatory as regulatory frameworks mature across jurisdictions.

---

## CSA Resource Alignment

The Cloud Security Alliance has published a substantial body of work directly applicable to the governance challenges raised by the NIST AI Agent Standards Initiative and the emerging regulatory landscape. Security teams should treat these resources as the primary implementation toolkit while government frameworks are finalized.

The AI Controls Matrix (AICM) v1.0 provides 243 controls across 18 security domains designed specifically for AI systems, extending the proven Cloud Controls Matrix v4.0 with AI-specific requirements including a dedicated Model Security domain [10]. The AICM's Shared Security Responsibility Model explicitly delineates control ownership across the five AI supply chain actors — cloud service providers, model providers, orchestrated service providers, application providers, and AI customers — making it directly applicable to the multi-party architectures typical of agent deployments. Organizations mapping their controls posture against NIST AI RMF or preparing for EU AI Act high-risk compliance will find the AICM's framework mappings a valuable starting point.

CSA's Agentic AI Red Teaming Guide addresses twelve threat categories for autonomous agents in depth, with step-by-step testing procedures, tool recommendations, and example prompts for conducting security assessments before and after deployment. The guide's coverage of authorization hijacking, hallucination exploitation, multi-agent exploitation, and agent untraceability directly maps to the threat categories identified in CAISI's RFI, providing a practical testing methodology aligned with the direction of government guidance [11].

The AI Organizational Responsibilities series — covering core security responsibilities, governance and risk management, and cultural aspects of AI accountability — addresses the human oversight and accountability structures that both NISTIR 8596 and the Singapore IMDA framework identify as essential governance controls [12]. The MAESTRO framework for agentic AI threat modeling provides threat modeling methodology specifically designed for the multi-component architectures of agent systems, filling a gap that general-purpose threat modeling approaches do not adequately address.

Organizations pursuing formal assurance should note that CSA's STAR for AI certification program provides structured pathways incorporating the AICM, with three distinct certification tracks based on organizations' existing ISO 27001 or ISO 42001 certification status [10]. This certification program provides the kind of third-party attestation that high-risk AI Act obligations and enterprise procurement programs are increasingly expected to require.

---

# References

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

[2] NIST CAISI, "CAISI Issues Request for Information About Securing AI Agent Systems," nist.gov, January 12, 2026. https://www.nist.gov/news-events/news/2026/01/caisi-issues-request-information-about-securing-ai-agent-systems

[3] Federal Register, "Request for Information Regarding Security Considerations for Artificial Intelligence Agents," Document 2026-00206, January 8, 2026. https://www.federalregister.gov/documents/2026/01/08/2026-00206/request-for-information-regarding-security-considerations-for-artificial-intelligence-agents

[4] NIST, "Draft NIST Guidelines Rethink Cybersecurity for the AI Era" (NISTIR 8596, preliminary draft – Cybersecurity Framework Profile for Artificial Intelligence), nist.gov, December 16, 2025. https://www.nist.gov/news-events/news/2025/12/draft-nist-guidelines-rethink-cybersecurity-ai-era

[5] European Parliament IMCO/LIBE Committees, Report A10-0073/2026 on the EU Digital Omnibus proposal (AI Act delay amendment), adopted 101–9–8, March 18–19, 2026. https://www.europarl.europa.eu/doceo/document/A-10-2026-0073_EN.html

[6] European Commission, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act)," Official Journal of the European Union, August 1, 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[7] Infocomm Media Development Authority (IMDA), Singapore, "New Model AI Governance Framework for Agentic AI," press release, January 22, 2026. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2026/new-model-ai-governance-framework-for-agentic-ai

[8] CISA, "News & Events," cisa.gov, accessed March 22, 2026. https://www.cisa.gov/news-events (Note: site advisory noted lapse in federal funding.)

[9] CISA, NSA AI Security Center, FBI, and international partner agencies, "Principles for Secure Integration of Artificial Intelligence in Operational Technology," cisa.gov, December 3, 2025. https://www.cisa.gov/resources-tools/resources/principles-secure-integration-artificial-intelligence-

operational-technology

[10] Cloud Security Alliance, "AI Controls Matrix (AICM) v1.0 and STAR for AI Certification," cloudsecurityalliance.org. https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

[11] Cloud Security Alliance, "Agentic AI Red Teaming Guide," AI Organizational Responsibilities Working Group, 2025. https://cloudsecurityalliance.org/research/topics/artificial-intelligence

[12] Cloud Security Alliance, "AI Organizational Responsibilities: Governance, Risk Management, Compliance and Cultural Aspects," 2025. https://cloudsecurityalliance.org/research/topics/artificial-intelligence

 