NIST AI Agent Security: Red-Teaming Guidance and Enterprise Compliance

The AI Agent Standards Initiative, Adversarial Testing Findings, and Compliance Implications for 2026

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-31

Key Takeaways

- NIST's Center for AI Standards and Innovation (CAISI) formally launched the **AI Agent Standards Initiative** on February 17, 2026 [1], establishing a three-pillar program to standardize agent security, interoperability, and identity — representing NIST's most explicit and comprehensive treatment of agentic AI as a distinct standardization priority, and the first time NIST has established a dedicated organizational initiative around agent security as a category.
- NIST's own red-team research found that novel attack techniques targeting AI agents achieved an **81% task-hijacking success rate**, compared to 11% for the strongest known baseline attacks — a result suggesting that agent-specific offensive research can dramatically outperform defenses calibrated to known attack taxonomies [2].
- The March 2025 update to **NIST AI 100-2** (Adversarial Machine Learning Taxonomy) extended NIST's adversarial ML attack taxonomy to cover autonomous AI agent vulnerabilities for the first time, including indirect prompt injection, agent memory poisoning, and supply chain attacks on agent tools [3].
- The NCCoE's February 2026 concept paper proposes a demonstration project for AI agent identity and authorization using OAuth 2.0, SPIFFE/SPIRE, and Model Context Protocol — offering enterprises an early preview of likely NIST technical guidance on agent identity ahead of formal special publications [4].
- **COSAiS** (Control Overlays for Securing AI Systems), NIST's forthcoming extension of SP 800-53 to AI use cases, will include dedicated overlays for single-agent and multi-agent deployments; once finalized, these overlays may provide the basis for future FedRAMP AI requirements — a compliance trajectory that organizations in regulated sectors should monitor [5].
- The NIST CAISI Request for Information on AI agent security (NIST-2025-0035) drew formal responses from the OpenID Foundation and a financial services industry coalition — BITS, the Bank Policy Institute, and the American Bankers Association — signaling that identity federation standards and financial sector risk management are among the highest-priority topics that industry and standards bodies are raising in formal NIST engagement processes [6][7].

Background

The Agentic AI Inflection Point

NIST's foundational Artificial Intelligence Risk Management Framework (AI RMF 1.0, NIST AI 100-1) was published in January 2023, at a point when most production-scale enterprise AI deployments consisted of narrowly scoped predictive models and early-generation language model assistants [8]. The framework's Govern-Map-Measure-Manage structure provided strong guidance for those deployment patterns, but it was not designed for systems that autonomously plan multi-step tasks, delegate subtasks to subordinate agents, invoke external tools, persist context across sessions, and execute real-world actions — the defining characteristics of modern agentic AI systems.

The past two years have seen the rapid productization of autonomous AI agents across enterprise functions: coding assistants that execute and test their own generated code, security operations agents that triage and respond to alerts, financial agents that read and act on real-time market data, and orchestration platforms that chain multiple specialized agents into end-to-end workflows. These systems present a fundamentally different risk profile from the static models the AI RMF was written to govern. They accumulate and act on information over time, operate across organizational trust boundaries, and can produce cascading real-world consequences from a single compromised instruction.

NIST's response to this shift has been a substantial body of supplementary guidance published across 2025 and early 2026. Taken together, this work constitutes an emerging AI agent security framework — not yet consolidated into a single revised document, but coherent in its threat model, methodology, and compliance trajectory.

NIST CAISI and the AI Agent Standards Initiative

The Center for AI Standards and Innovation (CAISI) within NIST's Information Technology Laboratory serves as the agency's coordinating body for AI standards development, both domestically and in international bodies such as ISO/IEC JTC 1/SC 42 and the International Telecommunication Union. On February 17, 2026, CAISI announced the formal **AI Agent Standards Initiative** — the most direct signal yet that NIST is treating agentic AI as a distinct and urgent standardization priority [1].

The initiative is organized around three pillars. The first facilitates industry-led development of agent standards and positions U.S. participants for leadership in international standards bodies, particularly on agent interoperability and security specifications [12]. The second pillar fosters community-led development and maintenance of open-source agent protocols, including the Model Context Protocol

ecosystem. The third advances fundamental research in AI agent security, with specific emphasis on agent identity — the problem of reliably distinguishing AI agents from human users within enterprise systems, and ensuring that agents can only act within explicitly delegated scopes of authority.

CAISI's January 2026 Request for Information (Federal Register docket NIST-2025-0035) preceded the initiative's formal launch and solicited public input on five core questions: the unique security threats posed by AI agents, methods for improving security during development and deployment, gaps in applying existing cybersecurity frameworks to agents, risk measurement techniques, and deployment environment interventions [6]. The comment period closed March 9, 2026. While NIST has not yet released a compiled analysis of submissions, early responses from the OpenID Foundation and the financial services coalition of BITS, the Bank Policy Institute, and the American Bankers Association indicate that authorization architecture and financial sector risk management are among the highest-priority topics industry submitted [6][7].

Security Analysis

NIST's Red-Teaming Research: What Empirical Testing Found

A particularly operationally significant finding in NIST's recent agent security body of work comes from CAISI's red-teaming research published in January 2025 and conducted in collaboration with the UK AI Security Institute. The research used the open-source AgentDojo evaluation framework — developed at ETH Zurich and providing 97 injection tasks across 629 test cases — and NIST's own enhancements built on the UK AISI's Inspect evaluation platform [2].

The headline result is stark: when red teamers developed novel attack techniques tailored to the specific behavioral patterns of LLM-backed agents — rather than relying on known baseline attack patterns — task-hijacking success rates rose from 11% to 81% [2]. This is not a marginal improvement; it represents a qualitative shift in what attackers can achieve once they invest in agent-specific offensive research. The implication is that defenses calibrated against known attack taxonomies are likely to substantially underestimate the real attack surface.

NIST drew four operational conclusions from this research. First, agent evaluation frameworks require continuous iteration as both models and attack techniques evolve — a single benchmark score is not a reliable steady-state security indicator. Second, because novel techniques so substantially outperform known baselines, red team exercises that rely entirely on existing tooling and playbooks provide a false sense of assurance. Third, aggregate success rate statistics mask wide per-task variation; different injection tasks present substantially different risk profiles, and risk assessments must account for task-

specific outcomes rather than average performance. Fourth, because LLM agents produce non-deterministic outputs, security testing must model multi-attempt scenarios rather than single-shot evaluations — the research found that with 25 repeated attempts per task, average attack success rates climbed from 57% to 80% [2].

These findings align with and extend the empirical picture developed by NIST's ARIA program (Assessing Risks and Impacts of AI, AI 700-2), published in November 2025 [9]. The ARIA 0.1 pilot involved approximately 51 red teamers and 508 testing sessions across seven submitted AI applications, and introduced the Contextual Robustness Index (CoRIx) as a metric measuring the degree to which an AI application's output meets requirements for its intended use context [9]. ARIA establishes a three-tier evaluation hierarchy — model testing, adversarial red teaming, and field testing — that enterprises can use as a template for structuring internal AI security validation programs.

The Adversarial ML Taxonomy: Agent-Specific Threats

NIST's March 2025 update to AI 100-2 (Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2 E2025) provides the authoritative classification framework for attack categories relevant to AI agent deployments [3]. The 2025 edition extended coverage from the 2023 original in several significant ways.

A particularly consequential addition is the dedicated treatment of autonomous AI agent vulnerabilities, which were absent from the 2023 edition. The 2025 taxonomy explicitly covers indirect prompt injection — attacks in which a threat actor plants adversarial instructions not in a direct user query, but in data sources that an agent will later read and act upon: web pages, documents, database records, email messages, calendar entries, or tool outputs. Because agents are designed to act on information they retrieve from the environment, indirect prompt injection attacks can achieve code execution or data exfiltration without any access to the user's direct interaction with the system, in some configurations at high success rates. This attack class is structurally resistant to the input validation controls that defend against direct prompt injection.

The 2025 taxonomy also expands coverage of poisoning attacks relevant to agent architectures. Agents that accumulate knowledge over time through retrieval-augmented generation, long-term memory modules, or repeated interaction with external data sources are vulnerable to gradual corruption of that knowledge base — enabling an attacker who can influence any data source the agent reads to condition the agent's future behavior. The taxonomy provides mitigation guidance for each category, though it acknowledges that agent-specific mitigations remain less mature than those for the classic evasion and poisoning attacks against supervised learning models [3].

Agent Identity: The Authorization Gap

The NCCoE concept paper published February 5, 2026, "Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization," identifies agent identity and authorization as a foundational gap in enterprise AI security architectures [4]. The document proposes a demonstration project to show how existing enterprise identity and authorization standards can be extended to AI agents without requiring entirely new infrastructure.

The NCCoE concept paper identifies two common anti-patterns in how enterprises currently authorize AI agent actions. Agents that inherit full user-level permissions acquire far more access than any individual task requires; agents operating under generic service account credentials cannot be audited back to specific user intent. Neither pattern is acceptable in environments subject to least-privilege requirements, audit obligations, or data residency constraints [4].

The NCCoE project proposes using OAuth 2.0 and its extensions (including Rich Authorization Requests, Pushed Authorization Requests, and Demonstrating Proof-of-Possession) as the authorization layer for agent-initiated actions, binding agent sessions explicitly to the identity and delegated scope of the human user they serve [4]. SPIFFE/SPIRE provides workload identity for agents as computational entities, enabling cryptographic attestation of agent identity independent of human credential chains. The Model Context Protocol appears in the concept paper as the agent communication layer that would carry these identity assertions between agents and the tools they invoke. The concept paper's authorization model provides an early preview of the technical direction NIST is likely to take in forthcoming special publications.

COSAiS: Bringing SP 800-53 to AI Agents

The Control Overlays for Securing AI Systems (COSAiS) project at NIST's Computer Security Resource Center represents the compliance pathway with the most immediate enterprise significance [5]. SP 800-53 is the primary control catalog for federal cybersecurity compliance: it underlies FedRAMP authorization, CMMC requirements, and numerous sector-specific frameworks. Organizations that provide cloud services to federal agencies, or that operate in regulated sectors with frameworks derived from SP 800-53, should treat the COSAiS discussion draft as early signal of requirements that may eventually affect their compliance posture — though no formal FedRAMP adoption of COSAiS has been announced.

COSAiS defines five AI deployment use cases, each of which will receive a dedicated control overlay: using and adapting generative AI assistants, fine-tuning predictive AI, deploying single-agent AI systems, deploying multi-agent AI systems, and developing AI systems [5]. The single-agent and multi-agent overlays are the most directly relevant to enterprise security teams. The project published an annotated

outline discussion draft in January 2026 for initial feedback; overlays for the agentic AI use cases are expected in subsequent drafts. The project was created in July 2025 and published its initial concept paper in August 2025, placing full publication on a timeline of late 2026 to 2027 — though the discussion draft stage already provides sufficient signal for enterprises to begin gap analysis.

The Cyber AI Profile (NIST IR 8596, preliminary draft published December 2025) offers a parallel compliance onramp for organizations whose primary framework is CSF 2.0 rather than SP 800-53 [10]. The profile maps CSF 2.0 Functions, Categories, and Subcategories to three AI security objectives — securing AI system components, conducting AI-enabled cyber defense, and countering AI-enabled cyberattacks. Organizations already operating mature CSF 2.0 programs can use this profile to extend existing compliance posture to cover AI systems without implementing a separate framework.

Agent-Specific Threats Synthesized from NIST Guidance

Drawing on the three primary NIST sources surveyed in this note — the Agent Standards Initiative announcement, the CAISI RFI, and the AI 100-2 E2025 taxonomy — six threat categories emerge as specifically relevant to agentic AI deployments and distinguish them from conventional AI security concerns. Autonomous behavior resulting in real-world actions requires human oversight mechanisms that do not exist for static model inference. Dynamic tool-switching defeats static policy enforcement because the attack surface expands and contracts with each agent session. Information retention over time enables adversaries to poison agent behavior through data sources that the agent will later act upon. Non-deterministic output defeats rule-based security controls calibrated to expected behavior ranges. Indirect prompt injection remains resistant to input validation because the malicious instruction arrives through legitimate data retrieval channels. Finally, specification gaming — where an agent optimizes for measurable outcomes in ways that violate the intent of its instructions — can produce harmful results even without adversarial inputs.

Recommendations

Immediate Actions

Enterprises should take stock of their current AI agent inventory with the same rigor applied to software asset management. Every production AI agent deployment — whether a commercial product, a vendor-provided service, or an internally built system — should be documented with its scope of tool access, the data sources it reads and writes, the human identities it acts on behalf of, and the authorization model governing its actions. This inventory is the prerequisite for any meaningful gap analysis against NIST's

emerging agent security guidance. Organizations that cannot enumerate what agents they have deployed are not in a position to assess or demonstrate compliance with forthcoming COSAiS overlays or any future FedRAMP requirements that may incorporate them.

Security teams should incorporate the NIST AI 100-2 E2025 adversarial ML taxonomy into red team exercise planning. Specifically, the taxonomy's coverage of indirect prompt injection and agent memory poisoning defines concrete test-case categories that existing penetration testing programs typically do not include. Red team exercises against AI agent systems should be scoped to test whether adversarial instructions planted in documents, email, calendar data, and retrieved web content can influence agent behavior — not merely whether the agent resists direct adversarial prompts. NIST's open-source AgentDojo-Inspect toolchain provides a structured starting point for teams that need an evaluation framework [2].

Short-Term Mitigations

Authorization architecture for AI agent systems should be reviewed against the least-privilege principles outlined in the NCCoE concept paper, even ahead of formal NIST publication [4]. Agents operating under full user-level permissions or generic service accounts represent a significant privilege escalation risk; an agent compromised through indirect prompt injection with broad system access can cause far greater harm than one constrained to a specific delegated scope. Where enterprise identity infrastructure supports OAuth 2.0 scoping, delegated authorization chains should be implemented to bind agent permissions to specific tasks and time windows. Organizations with mature identity and access management programs should treat the NCCoE concept paper as a preview of forthcoming NIST technical guidance and begin mapping their current agent access patterns against the authorization model it proposes.

Organizations should pilot the ARIA three-tier evaluation model — model testing, adversarial red teaming, and field testing — as an internal assessment structure for AI applications ahead of external regulatory requirements. The ARIA program's Contextual Robustness Index framework provides a structured way to document AI system fitness for purpose that will translate well into future compliance submission formats. NIST's finding that multi-attempt testing is necessary for accurate risk assessment of probabilistic LLM systems should be reflected in testing protocols; single-shot evaluations that achieve acceptable scores may not represent realistic attacker capability [9].

Strategic Considerations

Enterprise AI governance programs should be structured to accommodate the compliance trajectory that NIST's current body of work foreshadows. COSAiS will extend SP 800-53 controls to AI agent deployments; once finalized, these overlays may provide the basis for future FedRAMP AI requirements [5]. Organizations that may be on that trajectory should begin mapping their agent deployments to the emerging COSAiS use case categories now, treating the discussion draft as an early-stage preview of potential requirements rather than waiting for final publication.

The NIST AI Agent Standards Initiative's engagement with Model Context Protocol development signals that protocol-level security properties — agent authentication, tool call authorization, and audit logging at the MCP layer — will be part of formal standards as they mature [1]. Security architects designing agent platforms and integration layers should treat MCP security as a first-class requirement, not a deferred hardening task. Organizations that build agent infrastructure with identity and audit capabilities baked in will find it substantially easier to satisfy forthcoming compliance requirements than those that retrofit these properties onto existing deployments.

Monitoring NIST's forthcoming release of RFI response analyses (docket NIST-2025-0035, comment period closed March 9, 2026) and the NCCoE concept paper feedback cycle (open through April 2, 2026) will provide early signal on how NIST intends to prioritize specific threat categories and technical approaches in subsequent draft publications. Submitting public comments to open NIST processes is a concrete way for organizations to ensure that operational constraints and implementation realities are reflected in final guidance.

CSA Resource Alignment

The NIST AI agent security body of work surveyed in this note maps directly onto several existing CSA frameworks and publications, making CSA guidance a practical implementation complement to NIST's policy and compliance infrastructure.

The **CSA Agentic AI Red Teaming Guide** [13] (2025) provides operationally detailed guidance for translating NIST's red-teaming recommendations into executable test plans. The Guide's taxonomy of 12 threat categories — including Agent Authorization and Control Hijacking, Agent Knowledge Base Poisoning, Agent Memory and Context Manipulation, and Multi-Agent Orchestration Exploitation — maps closely to the attack categories now formalized in NIST AI 100-2 E2025 and the threat inventory developed through the CAISI RFI process. Security teams conducting AI agent red team exercises

should cross-reference both documents: NIST provides the authoritative taxonomy and quantitative benchmark data, while the CSA Guide provides the procedural depth and example test cases needed for execution.

MAESTRO (CSA's agentic AI threat modeling framework) provides the architectural decomposition needed to apply NIST's threat categories to specific agent designs. NIST's emerging guidance operates at the policy and compliance level; MAESTRO provides the layer-by-layer threat model that allows security architects to reason about where specific NIST-identified threats manifest in a given agent's architecture — at the model layer, the orchestration layer, the tool interface layer, or the data access layer.

The **AI Controls Matrix (AICM) v1.0** [14], with its 243 security controls across 18 domains organized around a shared security responsibility model, provides an existing control framework against which COSAiS overlays can be mapped once finalized. The AICM's separation of responsibilities across Model Providers, Orchestrated Service Providers, Application Providers, Cloud Service Providers, and AI Customers aligns with NIST's multi-party agentic deployment model and provides a control attribution structure that organizations can use to demonstrate compliance ownership as formal requirements mature. Because AICM is a superset of CCM [14], organizations using AICM for AI governance are also maintaining posture against CCM-based compliance obligations.

The **STAR for AI** program provides the assessment and certification mechanism through which organizations can document their AI agent security posture against the AICM and demonstrate compliance readiness to auditors and customers. As COSAiS overlays and any future FedRAMP AI requirements mature, STAR for AI Level 2 assessments — grounded in ISO/IEC 42001 and the AI-CAIQ — will provide the third-party attestation pathway that enterprises in regulated sectors will need. Organizations preparing for forthcoming NIST-derived compliance requirements should treat STAR for AI readiness as a parallel workstream to their gap analysis against COSAiS.

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. NIST CAISI Technical Staff, "Strengthening AI Agent Hijacking Evaluations," NIST Technical Blog, January 17, 2025 (updated December 19, 2025). https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

3. NIST, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," NIST AI 100-2 E2025, March 24, 2025. https://doi.org/10.6028/NIST.AI.100-2e2025

4. NIST NCCoE / ITL, "Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization," NCCoE Concept Paper (Initial Preliminary Draft), February 5, 2026. https://csrc.nist.gov/pubs/other/2026/02/05/accelerating-the-adoption-of-software-and-ai-agent/ipd

5. NIST CSRC, "Control Overlays for Securing AI Systems (COSAiS)," CSRC Project Page, created July 10, 2025; Annotated Outline Discussion Draft published January 8, 2026. https://csrc.nist.gov/projects/cosais

6. NIST CAISI, "CAISI Issues Request for Information About Securing AI Agent Systems," NIST News, January 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

7. OpenID Foundation, "OIDF Responds to NIST on AI Agent Security," OpenID Foundation Blog, March 11, 2026. https://openid.net/oidf-responds-to-nist-on-ai-agent-security/; BITS / Bank Policy Institute / American Bankers Association, "Joint Letter on AI Security Considerations RFI," March 9, 2026. https://www.aba.com/advocacy/policy-analysis/joint-letter-on-ai-security-considerations-rfi

8. NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, January 26, 2023. https://www.nist.gov/itl/ai-risk-management-framework

9. NIST, "Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Report," NIST AI 700-2, November 2025. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.700-2.pdf

10. NIST, "Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile," NIST IR 8596 (Initial Preliminary Draft), December 16, 2025. https://csrc.nist.gov/pubs/ir/8596/iprd

11. Cloud Security Alliance, "Agentic AI Red Teaming Guide," AI Organizational Responsibilities Working Group (Ken Huang, lead), 2025. https://cloudsecurityalliance.org

12. Cloud Security Alliance, "Introductory Guidance to the AI Controls Matrix (AICM) v1.0," 2025. https://cloudsecurityalliance.org

---

Further Reading

The following NIST publications provide broader context for the agent security guidance surveyed in this note. They are not directly cited in the analysis above but inform the policy environment in which CAISI's work is situated.

- NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST AI 600-1, July 26, 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf – Addresses agentic AI patterns including multi-agent systems and tool-using architectures within the GenAI context; provides complementary guidance to the CAISI initiative's dedicated agent security program.

- NIST, "A Plan for Global Engagement on AI Standards," NIST AI 100-5 E2025, April 2025. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5e2025.pdf – Details NIST's international AI standards engagement strategy, relevant to CAISI's first pillar on positioning U.S. participants for leadership in international standards bodies.

