Agentic AI Governance: NIST Standards for Autonomous Systems

A Practitioner's Guide to Emerging Federal Frameworks for AI Agent Risk Management

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-22

Table of Contents

Executive Summary

The rapid proliferation of agentic AI systems—software architectures in which large language models plan multi-step tasks, invoke external tools, and execute consequential actions with minimal human supervision—has outpaced the governance frameworks that organizations use to manage AI risk. The foundational NIST AI Risk Management Framework (AI RMF 1.0, January 2023) was developed before agentic architectures emerged as a mainstream deployment pattern. Its four-function structure—Govern, Map, Measure, Manage—provides a well-established governance architecture, but was not designed to address the distinctive hazards that arise when AI systems can autonomously chain decisions, escalate privileges, interact with external services, and produce real-world effects at machine speed.

The standards ecosystem is catching up. NIST's Center for AI Standards and Innovation (CAISI) launched the AI Agent Standards Initiative in February 2026, signaling that purpose-built governance guidance for autonomous systems is now a federal priority. An updated edition of NIST AI 100-2 (March 2025) explicitly names AI agents as a threat surface for the first time. A preliminary draft of NIST IR 8596 published in December 2025 maps cybersecurity framework functions to AI-specific risks, including agentic threats. The NIST National Cybersecurity Center of Excellence (NCCoE) released a concept paper in February 2026 on software and AI agent identity and authorization. Together these workstreams represent NIST's most direct institutional response to date to the agent governance gap.

This whitepaper examines the current and emerging NIST standards landscape for agentic AI governance, situates it within the broader international regulatory context, and connects it to CSA's own frameworks—particularly the AI Controls Matrix (AICM), the MAESTRO threat modeling framework, and CSA's prior research on agentic AI identity and secure design. Security architects, governance professionals, and risk officers who are integrating AI agents into enterprise environments will find in this document a structured roadmap for aligning their programs to the evolving requirements of 2026 and beyond.

Introduction: The Agent Governance Gap

This whitepaper was prepared by the CSA AI Safety Initiative. Several CSA frameworks—MAESTRO, AICM, and CSA agentic AI research publications—are cited as complementary resources throughout the document. Readers should evaluate these alongside other available frameworks when building their governance programs.

Reports from major technology vendors and industry analysts suggest that organizations across financial services, healthcare, technology, and critical infrastructure are deploying autonomous AI agents to accelerate operations and reduce the cost of knowledge work. As of early 2026, major cloud providers, enterprise software vendors, and independent software developers have released agent frameworks, agent orchestration platforms, and agent-enabled applications that can browse the web, write and execute code, query databases, send email, manage files, interact with APIs, and coordinate fleets of subordinate agents—all in response to a high-level human instruction.

The governance frameworks most organizations rely on were designed for a different kind of AI deployment: a model receives an input, generates an output, and a human decides what to do with that output. Agentic systems break this pattern in ways that are consequential for risk management. An agent may execute dozens or hundreds of tool calls to complete a single task. It may spawn sub-agents with their own tool access. It may interact with external services in ways that are difficult to audit in real time. It may produce effects that are difficult or impossible to reverse—a sent email, a deleted file, a completed financial transaction, an executed API call that triggers downstream business processes. And when multiple agents interact in a pipeline, failures and adversarial manipulations can propagate in ways that no single-agent governance model anticipates.

This governance gap—the mismatch between the capabilities of deployed agentic systems and the frameworks organizations use to govern them—is the central problem this whitepaper addresses. It does not argue that existing frameworks are irrelevant. The NIST AI RMF, ISO/IEC 42001, the EU AI Act, and CSA's AICM all provide essential infrastructure. But each requires thoughtful extension to address the specific hazards of autonomous, tool-using, multi-agent systems. Understanding what those extensions look like, and how the emerging NIST AI Agent Standards Initiative is beginning to provide them, is the work of governance professionals today.

The NIST AI Risk Management Framework: Foundation and Gaps

NIST AI 100-1, the AI Risk Management Framework, was published on January 26, 2023, as a voluntary, sector-agnostic standard for managing the risks of AI systems across their full lifecycle [1]. Its four core functions—Govern, Map, Measure, and Manage—provide a logical structure that organizations can adapt to their AI deployments regardless of industry, AI modality, or regulatory context. The framework emphasizes trustworthy AI attributes including validity and reliability, safety, security and resilience, explainability and interpretability, privacy, and fairness.

The Govern function establishes the organizational policies, accountability structures, and culture necessary for responsible AI development and use. Map focuses on identifying the context and conditions of AI system deployment, understanding stakeholder impacts, and characterizing risks. Measure directs organizations to develop quantitative and qualitative methods to analyze and assess identified risks. Manage addresses the prioritization and treatment of risks through plans, responses, and continuous monitoring. Supplementary resources—the AI RMF Playbook, sector-specific profiles, and crosswalk documents linking the RMF to ISO/IEC 42001, the EU AI Act, and other frameworks—extend the core document's utility for practitioners [2].

The framework's voluntary status is a deliberate design choice that reflects NIST's historical role in U.S. standards development: producing technically rigorous guidance that industry, academia, and government can adopt, adapt, and build upon without the compliance burden of regulation. As of early 2026, no U.S. federal regulation mandates AI RMF adoption, though several regulatory agencies reference it in guidance and proposals.

What the AI RMF 1.0 was not designed to address are the unique governance challenges of agentic AI. The document's risk categories—data quality, model performance, adversarial robustness, fairness, privacy—apply at the level of individual AI model inference. Agentic systems introduce additional risk dimensions that operate at the system level: the risk that an agent executes unintended actions based on a flawed plan, the risk that a prompt injection attack hijacks an agent's decision-making process, the risk that an agent escalates its own access privileges beyond what was authorized, the risk that multi-agent communication channels become vectors for cross-agent manipulation, and the risk that automated action chains produce harm at a speed and scale that human oversight cannot effectively monitor. These risks require governance interventions that go beyond the model-centric framing of the current RMF. The AI Agent Standards Initiative, discussed in the next section, is NIST's response to this need.

The NIST AI Agent Standards Initiative

On February 17, 2026, NIST's Center for AI Standards and Innovation announced the AI Agent Standards Initiative, a significant federal standards effort specifically targeting autonomous AI systems [3]. The initiative is organized around three pillars that reflect both NIST's technical mission and the current competitive context for AI standards leadership.

The first pillar focuses on industry-led standards development, with NIST facilitating U.S. participation and leadership in international standards bodies—particularly ISO/IEC—on questions of AI agent interoperability, security, and governance. This work is consequential because international standards for AI agents, once established, will shape procurement requirements, regulatory conformity frameworks, and technical

architecture choices for years. U.S. participation at the standards table during this formative period can help ensure that security and governance considerations are weighted alongside interoperability requirements in the resulting standards.

The second pillar focuses on open-source protocol development, with NIST fostering community-maintained protocols that enable agents to interoperate securely. The absence of standard protocols for agent-to-agent communication, for agent identity attestation, and for expressing agent authorization scopes has led to a proliferation of vendor-specific approaches that impede governance. Without standard protocols, organizations cannot consistently verify the identity of agents they interact with, cannot uniformly express what those agents are and are not permitted to do, and cannot effectively audit agent interactions across system boundaries.

The third pillar is research in AI agent security and identity, advancing the foundational technical knowledge necessary to enable trusted adoption. As of March 2026, NIST had issued a Request for Information on AI agent security (comment period closed March 9, 2026), released a draft concept paper on agent identity and authorization through the NCCoE (comment period open through April 2, 2026), and announced benchmark evaluation workstreams with a comment period closing March 31, 2026. Sector-specific listening sessions are scheduled to begin in April 2026 [3].

The NCCoE Agent Identity and Authorization Concept Paper

Among the early outputs of the initiative, the NCCoE concept paper "Accelerating the Adoption of Software and AI Agent Identity and Authorization," released in February 2026, is the most immediately actionable for practitioners, addressing the concrete governance problem of agent identity and authorization [4]. This document confronts one of the most fundamental and underappreciated challenges in agentic AI deployment: how do we know who an agent is, what it is authorized to do, and how do we enforce those boundaries at runtime?

The concept paper identifies OAuth 2.0 and its extensions as a foundational technology for agent authorization, while acknowledging that current OAuth deployments were not designed with autonomous agents in mind. When a human user authenticates to an application, they provide credentials that establish identity; the resulting token scopes constrain what the application can do on their behalf. When an agent acts autonomously—spawning sub-agents, calling APIs, accessing data stores—the token issued to the originating human may be passed, delegated, or implicitly assumed in ways that expand access far beyond what was intended. The concept paper proposes policy-based access controls designed for agent delegation chains, mechanisms for expressing fine-grained authorization constraints on agent tool access, and approaches to prompt injection mitigation at the authorization layer [4].

The identity dimension of the problem is equally challenging. Enterprise identity infrastructure has historically been designed primarily for human users, with non-human identities—service accounts, API keys, automation tokens—typically managed through less mature practices relative to human identity management. Agentic systems change this calculus: an enterprise running dozens or hundreds of agents, each capable of acting autonomously across multiple systems, faces an identity management challenge that is qualitatively different from managing a set of static service accounts. The concept paper points toward decentralized identifier (DID) approaches and verifiable credentials as potential mechanisms for establishing portable, attestable agent identity, though it does not mandate specific technologies [4].

Adversarial Threats to Agentic Systems: NIST AI 100-2 E2025

A second critical 2025 development in the NIST AI standards landscape is the March 2025 publication of an updated edition of NIST AI 100-2, the adversarial machine learning taxonomy [5]. The original January 2024 edition established a classification framework for adversarial attacks—evasion, poisoning, and privacy attacks—focused primarily on predictive AI systems. The E2025 edition is the first NIST publication to explicitly name autonomous AI agents as a security threat surface.

The E2025 edition expands the attack taxonomy to cover generative AI systems and the architectures built on top of them. It introduces specific categories for prompt injection—the technique by which adversarial instructions embedded in external content (web pages, documents, tool outputs) hijack an agent's reasoning and cause it to take unintended actions—information leakage through user interaction patterns, and training data compromise affecting models that underpin agentic applications [5]. It also addresses AI supply chain security risks, recognizing that agentic systems are typically assembled from multiple components—base models, fine-tuned layers, tool integrations, orchestration frameworks, memory systems—each of which represents a potential compromise point.

For practitioners governing agentic deployments, NIST AI 100-2 E2025 provides the most complete authoritative taxonomy currently available for classifying adversarial risks against agent systems. Its attack categories map naturally to governance controls: prompt injection risks require controls on agent input validation and instruction authority hierarchies; supply chain risks require controls on model provenance and component integrity; privacy risks require controls on data minimization and output filtering. Organizations that have adopted the AI RMF's Measure function for adversarial risk assessment should update their assessment methodologies to incorporate the E2025 taxonomy.

The Cyber AI Profile: NIST IR 8596

In December 2025, NIST published an Initial Preliminary Draft of NIST IR 8596, the "Cybersecurity Framework Profile for Artificial Intelligence" (Cyber AI Profile) [6]. This document functions as a profile under NIST CSF 2.0 (published February 2024), mapping the CSF's six core functions—Govern, Identify, Protect, Detect, Respond, and Recover—to AI-specific cybersecurity risks. It was developed through a community of interest that engaged more than 6,500 contributors, reflecting broad industry and government participation.

The Cyber AI Profile addresses three overlapping problem spaces: securing AI systems themselves against attack and compromise; leveraging AI capabilities to enhance cyber defense operations; and understanding and countering AI-enabled cyberattacks. All three dimensions are relevant to agentic AI governance. Agentic systems must be secured as AI systems (the first dimension); they are increasingly being deployed for defensive cyber operations (the second); and adversaries are deploying agentic systems to conduct attacks at scale (the third).

For organizations that have already aligned their cybersecurity programs to CSF 2.0, the Cyber AI Profile provides a compatible integration path for AI-specific security requirements, using the same functional language as CSF 2.0. Rather than maintaining separate governance tracks for cybersecurity and AI risk management, the profile enables a unified framework in which AI security requirements are expressed alongside broader cybersecurity obligations. The comment period for the initial draft closed January 30, 2026; a revised draft is expected later in 2026 [6].

The Policy Context: U.S. Executive Action and NIST's Mandate

Understanding the current NIST AI standards landscape requires understanding the policy environment in which it operates. The Biden administration's Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (October 2023), directed NIST to develop a generative AI-specific companion to the AI RMF, resulting in NIST AI 600-1 (July 2024) [7]. EO 14110 also directed updates to the Secure Software Development Framework to address AI system development.

EO 14110 was rescinded on January 23, 2025, by the incoming Trump administration through Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence" [8]. EO 14179 directed federal agencies to review and revise policies issued under EO 14110 that were inconsistent with the new

administration's priorities. The America's AI Action Plan, released July 31, 2025, in response to EO 14179, directed NIST to revise the AI RMF's framing—removing certain equity and social impact language—but did not abolish the framework or NIST's broader AI standards mission [9].

Available evidence suggests the practical effect on NIST's core technical publications has been modest. The AI RMF's four-function governance architecture remains intact, and CAISI's agent standards work proceeded on schedule. However, the AI Action Plan did direct revisions to the RMF's framing, and the longer-term effect on NIST's priorities remains to be seen. For organizations building governance programs, the NIST AI standards ecosystem remains a reliable reference point, because those publications reflect broad technical consensus and community engagement regardless of administrative framing changes. The core voluntary guidance—AI RMF, AI 100-2, AI 600-1, IR 8596, and the agent standards workstreams—retains authoritative standing among practitioners.

A December 2025 executive order, "Ensuring a National Policy Framework for Artificial Intelligence," formalized administration interest in preempting conflicting state AI legislation, establishing a DOJ AI Litigation Task Force and directing FTC, FCC, and Commerce Department actions targeting state AI laws [21]. This development may eventually increase the practical authority of NIST guidance as a national baseline, though the longer-term effect on federal-state AI governance dynamics remains uncertain.

International Complementary Frameworks

ISO/IEC 42001: The Certifiable AI Management System

ISO/IEC 42001, published in December 2023 by ISO and the International Electrotechnical Commission, establishes requirements for an Artificial Intelligence Management System (AIMS)—a structured organizational framework for responsible AI governance analogous in design to ISO 27001 for information security [10]. Unlike the NIST AI RMF, ISO 42001 is a certifiable standard: organizations can demonstrate conformance through third-party audit and receive formal certification.

NIST has published an official crosswalk document mapping the AI RMF's functions and subcategories to ISO 42001 controls, enabling organizations to implement both frameworks in an integrated manner [11]. The practical relationship between them is complementary: the AI RMF provides a risk-based analytical model for identifying, assessing, and managing AI risks, while ISO 42001 provides the management system infrastructure—policies, organizational roles, continual improvement processes—necessary to operationalize those risk decisions consistently across the enterprise. ISO 42001 certification has begun to appear as a third-party attestation mechanism in enterprise procurement discussions and regulatory positioning, particularly for organizations preparing for EU AI Act conformity assessments.

ISO 42001 does not specifically address agentic AI systems, and the same gap analysis that applies to the NIST AI RMF applies here: the standard's controls were designed for AI systems in which human-AI interaction patterns are relatively well-defined. Organizations implementing ISO 42001 for governance programs that include agentic systems will need to develop supplementary controls addressing agent-specific risks—tool authorization, delegation chain integrity, prompt injection, and emergent multi-agent behavior—within the standard's management system structure.

CISA Guidance: From Models to Agents

The Cybersecurity and Infrastructure Security Agency has issued a progression of AI security guidance that increasingly addresses autonomous systems. The April 2024 joint publication "Deploying AI Systems Securely," co-issued with NSA, FBI, and international partner agencies, provided baseline guidance on confidentiality, integrity, and availability for AI deployments [12]. In May 2025, NSA, CISA, and FBI published joint AI data security guidance addressing data risks through the AI system lifecycle [22].

The December 2025 joint guidance "Principles for the Secure Integration of Artificial Intelligence in Operational Technology," issued with partners from Australia, Canada, Germany, the Netherlands, New Zealand, the United Kingdom, NSA, and FBI, marks the clearest federal acknowledgment to date that AI agents constitute a distinct category requiring dedicated governance attention [13]. The document explicitly defines AI agents as "a type of software that can process data, perform decision-making capabilities, and initiate autonomous actions using AI and ML models," and applies its four governance principles—understand AI, assess AI use in OT environments, establish AI governance, and manage AI risks—specifically to this category alongside predictive ML systems and LLM-based tools [13]. While the document's primary audience is critical infrastructure OT operators, its governance principles apply broadly to any organization deploying autonomous AI systems with access to consequential systems.

The EU AI Act and Agentic Systems

The European Union's AI Act, which entered into force August 1, 2024, creates the most comprehensive binding legal framework currently applicable to AI systems including autonomous agents, though the majority of its substantive requirements apply to EU-based deployments or entities placing AI systems on the EU market [14]. Understanding its provisions is essential for any global organization's AI governance program.

The Act's definition of an in-scope AI system—"machine-based systems designed to operate with varying levels of autonomy"—encompasses autonomous AI agents by design [14]. Governance requirements for agentic systems fall primarily across two regulatory tracks. The first track, governing general-purpose AI (GPAI) models, became operative on August 2, 2025, affecting the foundation model providers whose systems underpin most current agentic deployments. GPAI providers are now obligated to maintain

technical documentation, comply with copyright law, conduct model evaluations and adversarial testing, track and report serious incidents, and implement cybersecurity protections [14]. These obligations flow through the supply chain to organizations that deploy agents built on GPAI models, because deployment decisions affect how foundation model capabilities are used in ways that can trigger or avoid regulatory exposure.

The second track, governing high-risk AI systems, becomes fully operative on August 2, 2026. Agentic systems deployed in high-risk contexts—healthcare, critical infrastructure, law enforcement, employment and HR decisions, education, financial services—will face mandatory risk management systems, data governance obligations, logging and human oversight requirements, conformity assessments, and CE marking before EU market placement [14]. Organizations planning agent deployments in these sectors should treat the August 2026 deadline as an immediate governance priority, given the time required to complete conformity assessments and establish compliant system documentation.

The European Commission published draft guidelines for GPAI providers on July 18, 2025, clarifying the threshold for "systemic risk" designation at 10^25 floating point operations of training compute—a level that, as of early 2026, is publicly documented to have been reached by only a handful of frontier models from major AI developers—and specifying additional obligations for providers of models that cross that threshold [14].

CSA Framework Alignment

MAESTRO: Threat Modeling for Agentic AI

The CSA MAESTRO framework provides a purpose-built threat modeling methodology for agentic AI systems that complements the governance frameworks discussed in this whitepaper. Where the NIST AI RMF establishes organizational governance functions and ISO 42001 provides management system infrastructure, MAESTRO provides the threat model that identifies what agents are at risk from and how those risks manifest in multi-agent architectures. Security architects implementing agentic AI governance programs should treat MAESTRO as the primary threat identification tool, mapping its threat categories to controls from the AI Controls Matrix and governance processes from the AI RMF [15].

The AI Controls Matrix and NIST Alignment

CSA's AI Controls Matrix (AICM) v1.0.1 provides 243 security controls across 18 domains, structured around a five-role shared responsibility model that distributes accountability among AI providers, platform operators, application developers, security teams, and end users [16]. A CSA-published crosswalk maps all 243 AICM controls to NIST AI 600-1, enabling organizations to implement the AICM as an operational control layer

beneath the AI RMF's risk management functions [17]. Organizations that have adopted the AI RMF for risk governance can use the AICM crosswalk to translate risk findings into specific technical and procedural controls with defined ownership.

The AICM's 18 control domains include several that are directly relevant to agentic AI governance gaps: access control for AI systems, data governance and lifecycle management, model security and integrity, supply chain risk management, and incident detection and response. These domains align with the control categories identified in the NIST AI Agent Standards Initiative's NCCoE concept paper on agent identity and authorization, offering a productive integration path between the NIST and CSA control frameworks [4][16].

Agentic AI Identity and Zero Trust

CSA's research publication "Agentic AI Identity and Access Management: A New Approach" (August 2025) addresses the identity gap in current agent governance frameworks in detail [18]. The publication identifies specific inadequacies in OAuth 2.1 for agentic use cases—particularly the challenge of representing delegation chains in which a human authorizes an agent, the agent spawns sub-agents, and each sub-agent must operate within the scope of the original human authorization. It proposes decentralized identifiers (DIDs) and verifiable credentials as mechanisms for portable, attestable agent identity, and articulates an Agent Naming Service (ANS) concept for resolving agent identities across organizational boundaries [18].

This work is directly complementary to the NCCoE concept paper's focus on agent identity and authorization. Organizations evaluating both documents will find that CSA's publication provides greater technical depth on cryptographic identity mechanisms, while the NCCoE paper provides more detailed guidance on OAuth 2.0 extension patterns and policy-based access control architectures [4][18]. Together they represent the most complete available guidance on the agent identity problem prior to the publication of formal NIST standards in this area.

Zero trust principles, as articulated in NIST SP 800-207 [19], take on heightened urgency for agentic AI systems, where autonomous action at machine speed leaves limited time for after-the-fact detection and response. The zero trust mandate—never trust, always verify, enforce least privilege—becomes especially important when autonomous agents are capable of taking consequential actions across multiple systems. Every agent-to-resource interaction, every agent-to-agent communication, and every tool invocation should be subject to explicit authorization verification rather than implicit trust derived from network position or initial credential issuance. Governance programs should map zero trust enforcement points to agent interaction patterns and verify that existing IAM infrastructure is capable of handling the volume and velocity of authorization decisions that autonomous agent deployments require.

Secure Agentic System Design

CSA's "Secure Agentic System Design: A Trait-Based Approach" (August 2025) provides complementary guidance for the technical implementation dimension of agentic AI governance [20]. The publication identifies seven trait categories for secure agentic system design—including observability, least privilege, fail-safe defaults, defense in depth, and separation of privilege—and analyzes the governance implications of centralized versus decentralized agent control architectures. Organizations designing governance programs should use this framework alongside MAESTRO for threat modeling, ensuring that governance controls address not only what can go wrong in agent systems but how system design choices create or mitigate governance challenges.

Governance Implementation Guidance

Mapping the Current Landscape to Immediate Action

The standards landscape for agentic AI governance is actively forming rather than settled. Organizations should treat the period from early 2026 through 2027 as a **governance investment window**—a period in which programs built on currently available guidance will be validated and refined as NIST finalizes the AI Agent Standards Initiative workstreams, the Cyber AI Profile, and potential updates to the AI RMF itself. Standards development timelines are subject to change based on comment volume, resource priorities, and policy shifts; organizations should build governance programs around sound principles rather than specific anticipated publication dates.

Within this context, several actions are appropriate for governance programs in 2026. Organizations should conduct an inventory of deployed and planned agentic AI systems, documenting their tool access, data access, authorization patterns, and interaction with external services. This inventory is foundational to any subsequent governance activity: organizations cannot govern systems they have not identified.

The NIST AI RMF's Map function—focused on identifying the context, conditions, and risks of AI system deployment—provides the appropriate starting point for this inventory work. Applied to agentic systems, the Map function should encompass not only the agent's base model and training data but its tool integration surface, its authorization scope, its interaction with other agents, and the downstream systems whose state it can affect. Risk identification should incorporate the attack taxonomy from NIST AI 100-2 E2025, with particular attention to prompt injection, supply chain compromise, and information leakage risks [5].

Authorization Controls for Autonomous Agents

The identity and authorization gaps identified in the NCCoE concept paper and CSA's agentic AI identity research should be addressed through an immediate review of how current IAM infrastructure handles non-human agent identities [4][18]. Specifically, organizations should verify that agents operating in production environments have defined identity credentials—not merely inherited session tokens from human users—and that those credentials are associated with explicit authorization scopes that reflect the minimum access necessary for the agent's designated function. Authorization policies should define not only what resources an agent can access but what actions it can take on those resources, what other agents it can delegate to, and under what conditions its actions should be escalated to human review.

Where current IAM infrastructure does not support fine-grained agent authorization—for example, because it was designed for human-centric access patterns—organizations should develop compensating controls, such as agent-specific API gateway policies, tool invocation logging and alerting, and rate limits on agent action frequency.

Human Oversight Integration

Governance frameworks for agentic AI must address the human oversight challenge explicitly. The AI RMF's Manage function envisions risk management plans that include monitoring, response, and recovery activities; for agentic systems, these activities require oversight mechanisms that can operate at the speed and scale of autonomous action. Traditional audit log review is insufficient when agents can take hundreds of consequential actions per minute.

Effective human oversight for agentic systems requires interrupt conditions—predefined thresholds at which agent execution is paused and human review is required—combined with real-time monitoring infrastructure capable of detecting anomalous action patterns. Governance programs should define interrupt conditions as part of agent deployment governance, not as an afterthought. These conditions should include actions that exceed a defined impact threshold, actions that fall outside a defined authorization scope, actions that trigger anomaly detection alerts, and actions involving sensitive data categories.

Lifecycle Governance and Decommissioning

Agentic AI systems, like other software, require lifecycle governance from initial deployment through eventual decommissioning. The EU AI Act's high-risk system requirements make this a legal obligation for applicable deployments; for others, it is a governance best practice. Lifecycle governance should include version control for agent prompts and tool configurations, change management processes that require risk

assessment for significant changes to agent capability or authorization scope, and decommissioning procedures that address the disposition of agent credentials, conversation history, and any persistent state the agent has accumulated.

Supply chain governance deserves particular attention in the agentic context. An agent system is typically assembled from a base model, one or more fine-tuned layers, a retrieval or memory system, a tool integration layer, and an orchestration framework—each potentially supplied by a different vendor or open-source project. The supply chain attack surface is correspondingly large. Governance programs should maintain a software bill of materials (SBOM) for agent systems that includes model provenance, training data sourcing, and third-party tool and framework versions.

The Governance Horizon: What to Expect in 2026–2027

The NIST AI Agent Standards Initiative's sector-specific listening sessions are scheduled to begin in April 2026, with formal standards development activity expected to follow. The NCCoE agent identity and authorization concept paper comment period closes April 2, 2026, and a formal project announcement—which will initiate a multi-year collaborative standards development process—is anticipated shortly after. The Cyber AI Profile (IR 8596) is expected to advance from preliminary draft to revised draft in 2026, incorporating comments from the 6,500-contributor community of interest.

International standards activity will intensify in parallel. ISO/IEC working groups are actively developing standards for AI agent interoperability and security, with NIST participating through its CAISI mission. The EU AI Act's full implementation in August 2026 will create the first legally binding requirements applicable to high-risk agentic deployments, providing a practical forcing function for governance program completion in affected sectors.

Organizations that build governance foundations now—using the AI RMF, AICM, MAESTRO, and available NIST and CSA guidance on agent-specific risks—will be better positioned to adapt as formal standards emerge. The governance principles that will underpin future standards are already visible in the current landscape: identity and authorization for autonomous agents, prompt injection resistance, supply chain integrity, human oversight integration, and lifecycle accountability. Governance programs built around these principles will require updating as standards are finalized, but will not require rebuilding.

Conclusions and Recommendations

The governance landscape for autonomous AI agents is advancing rapidly but remains in formative stages. The February 2026 launch of NIST's AI Agent Standards Initiative represents a meaningful institutional commitment to federal standards leadership in this space, though the standards themselves will take several years to fully mature. Organizations deploying agentic AI systems today cannot wait for formal standards to govern those deployments; they must build governance programs on available foundations while maintaining the organizational capacity to adapt as the standards landscape develops.

The following recommendations are organized by governance function:

Govern. Establish formal accountability for agentic AI governance within the organization's AI risk management structure. Designate ownership for agent inventory, authorization policy, and incident response. Develop agentic AI governance policies that address the agent-specific risks identified in NIST AI 100-2 E2025—prompt injection, supply chain integrity, unauthorized delegation—as supplements to existing AI governance policies. Adopt ISO 42001 as the management system framework if third-party attestation is a near-term business requirement.

Map. Complete an inventory of all deployed and planned agentic AI systems, including their tool access, data access, authorization patterns, and downstream system interactions. Apply the MAESTRO threat modeling framework to each significant agent deployment. Identify the regulatory classification of agent deployments under the EU AI Act's high-risk categories and initiate conformity assessment planning for any deployment that falls within scope.

Measure. Implement the NIST AI 100-2 E2025 adversarial attack taxonomy as the basis for security assessment of agentic deployments. Establish monitoring metrics for agent behavior anomalies, tool invocation patterns, and authorization scope violations. Define interrupt conditions for each production agent deployment that will trigger human review.

Manage. Address the identity and authorization gap for agent credentials in current IAM infrastructure, using the NCCoE concept paper and CSA's agentic AI identity research as guidance. Establish SBOM practices for agent system components. Implement lifecycle governance processes including change management for agent prompt and configuration updates and decommissioning procedures for retired agents.

The governance frameworks analyzed in this whitepaper—NIST's AI RMF, AI 100-2, AI 600-1, IR 8596, the AI Agent Standards Initiative, ISO 42001, the EU AI Act, and CSA's MAESTRO, AICM, and agentic AI research—collectively provide a foundation adequate for initiating responsible agentic AI governance programs today, even as key standards remain in development. No single framework is complete, and organizations should expect to supplement current guidance as formal standards mature. The challenge is not the absence of frameworks but the discipline to apply them coherently to a technology that is changing faster than any

single framework can fully anticipate. Organizations that approach agentic AI governance with the same rigor they apply to other high-risk technology deployments will find that the available frameworks, properly applied, provide meaningful risk reduction even as the standards landscape continues to develop.

References

[1] National Institute of Standards and Technology. "AI Risk Management Framework (AI RMF 1.0)." NIST AI 100-1. January 26, 2023. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[2] National Institute of Standards and Technology. "AI RMF Playbook." NIST AI Risk Management Framework. Accessed March 2026. https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook

[3] National Institute of Standards and Technology, Center for AI Standards and Innovation. "Announcing the AI Agent Standards Initiative: Interoperable and Secure." February 17, 2026. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure

[4] National Cybersecurity Center of Excellence. "Accelerating the Adoption of Software and AI Agent Identity and Authorization." NIST NCCoE Concept Paper. February 2026. https://www.nccoe.nist.gov/sites/default/files/2026-02/accelerating-the-adoption-of-software-and-ai-agent-identity-and-authorization-concept-paper.pdf

[5] National Institute of Standards and Technology. "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations." NIST AI 100-2 (E2025 edition). March 24, 2025. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

[6] National Institute of Standards and Technology. "Cybersecurity Framework Profile for Artificial Intelligence." NIST IR 8596 (Initial Preliminary Draft). December 16, 2025. https://nvlpubs.nist.gov/nistpubs/ir/2025/NIST.IR.8596.iprd.pdf

[7] National Institute of Standards and Technology. "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

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

[9] Executive Office of the President. "Winning the Race: America's AI Action Plan." July 31, 2025. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[10] International Organization for Standardization and International Electrotechnical Commission. "ISO/IEC 42001:2023 — Artificial Intelligence — Management System." December 2023.

[11] National Institute of Standards and Technology. "NIST AI RMF to ISO/IEC 42001 Crosswalk." AI Resource Center. Accessed March 2026. https://airc.nist.gov/docs/NIST_AI_RMF_to_ISO_IEC_42001_Crosswalk.pdf

[12] Cybersecurity and Infrastructure Security Agency, National Security Agency, Federal Bureau of Investigation, et al. "Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems." April 2024. https://www.cisa.gov/news-events/alerts/2024/04/15/joint-guidance-deploying-ai-systems-securely

[13] Cybersecurity and Infrastructure Security Agency, Australian Signals Directorate, et al. "Principles for the Secure Integration of Artificial Intelligence in Operational Technology." December 3, 2025. https://www.cisa.gov/sites/default/files/2025-12/joint-guidance-principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology-508c.pdf

[14] European Parliament and Council of the European Union. "Regulation (EU) 2024/1689 on Artificial Intelligence (EU AI Act)." August 1, 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

[15] Cloud Security Alliance AI Safety Initiative. "MAESTRO: Threat Modeling Framework for Agentic AI Systems." Cloud Security Alliance. 2025.

[16] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.0.1." Cloud Security Alliance. August 14, 2025.

[17] Cloud Security Alliance. "AI Controls Matrix to NIST AI 600-1 Mapping." Cloud Security Alliance. July 7, 2025.

[18] Cloud Security Alliance AI Safety Initiative. "Agentic AI Identity and Access Management: A New Approach." Cloud Security Alliance. August 11, 2025.

[19] National Institute of Standards and Technology. "Zero Trust Architecture." NIST SP 800-207. August 2020. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf

[20] Cloud Security Alliance AI Safety Initiative. "Secure Agentic System Design: A Trait-Based Approach." Cloud Security Alliance. August 6, 2025.

[21] Executive Office of the President. "Ensuring a National Policy Framework for Artificial Intelligence." December 11, 2025. https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/

[22] National Security Agency, Cybersecurity and Infrastructure Security Agency, Federal Bureau of Investigation, et al. "AI Data Security: Best Practices for Securing Data Used to Train & Operate AI Systems." May 22, 2025. https://www.cisa.gov/resources-tools/resources/ai-data-security-best-practices-securing-data-used-train-operate-ai-systems