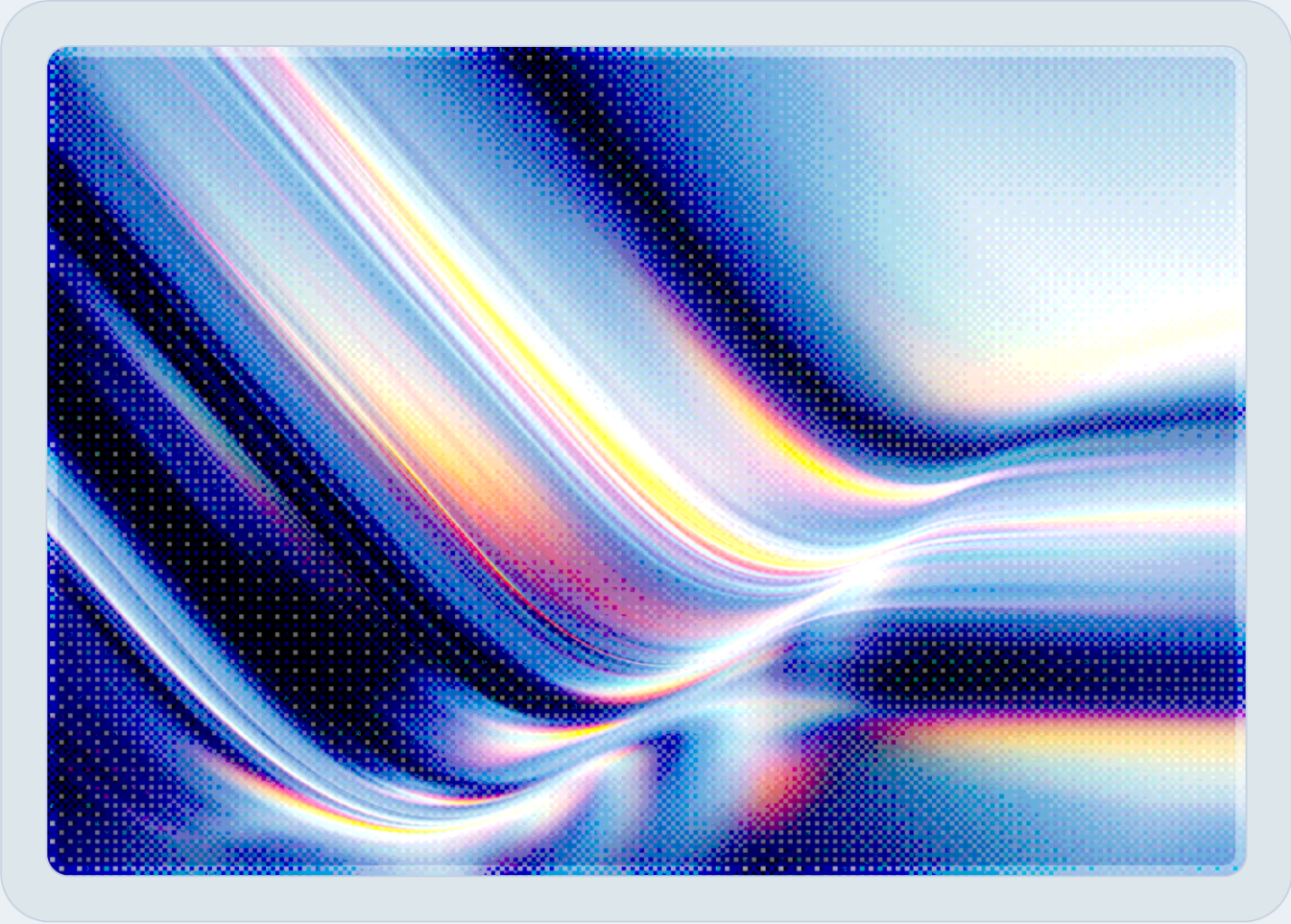
CSAI Foundation | Cloud Security Alliance

AI Agent Standards in 2026: NIST, CoSAI, AARM, and China

2026-05-18

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Four major governance efforts launched in early 2026 now define the contours of the AI agent security landscape: NIST's AI Agent Standards Initiative, the Coalition for Secure AI (CoSAI), the Autonomous Action Runtime Management (AARM) specification under the CSAI Foundation, and China's TC260 draft guidance on open-source AI agents.
- Agent identity and authentication have emerged as the single most critical and least-resolved security challenge across every framework—no widely adopted, purpose-built standard yet exists as of May 2026 for reliably authenticating autonomous agents or governing the delegations they carry.
- Fragmentation poses a near-term enterprise risk. Multinational organizations now face diverging compliance requirements across US-led frameworks and China's national standards, creating audit complexity before a single consensus standard has been finalized.
- The Model Context Protocol (MCP), implemented on more than 10,000 active public servers with over 97 million monthly SDK downloads as of late 2025 [1], has become the de facto agent-to-tool interface—expanding the attack surface faster than governance frameworks can track it.
- CSAI Foundation's AARM specification provides a detailed runtime enforcement architecture—including a context accumulator, policy engine, approval service, and telemetry exporter—while NIST's initiative and CoSAI's principles are still maturing toward implementable specifications.
- Security teams should treat agentic governance as an immediate operational priority, not a future compliance exercise. Inventory, identity lifecycle management, and runtime policy controls for agents are achievable today using existing frameworks.

Background

The emergence of AI agents—systems capable of planning and executing multi-step tasks autonomously across digital environments—has triggered a coordinated but still fragmented global response from standards bodies, industry coalitions, and national regulators. In the span of roughly four months between February and May 2026, four distinct governance efforts published major milestones, each approaching the problem from a different angle and with different institutional authority.

NIST's Center for AI Standards and Innovation (CAISI) announced the AI Agent Standards Initiative on February 17, 2026 [2]. The initiative is structured around three pillars: facilitating industry-led development of agent standards and US leadership in international standards bodies, fostering community-led open-source protocol development including the MCP ecosystem, and advancing fundamental research in AI agent security with particular emphasis on agent identity. Concurrent with the announcement, the National Cybersecurity Center of Excellence published a concept paper proposing to adapt existing identity and authorization frameworks for AI agents—an acknowledgment that current enterprise IAM infrastructure was not built for non-human identities operating at scale [3].

The Coalition for Secure AI (CoSAI), an OASIS Open Project with more than 45 sponsor organizations including Google, Microsoft, NVIDIA, IBM, and Meta, took a different path by publishing operational guidance ahead of formal standards processes. CoSAI's Workstream 4 released a paper on Agentic Identity and Access Management in March 2026, followed by Principles for Secure-by-Design Agentic Systems, which targets agent developers, adopters, and security engineers with practical implementation strategies [4][12]. CoSAI also released an extensive taxonomy for Model Context Protocol security [5], addressing the attack surface introduced by the protocol's rapid enterprise adoption.

The CSAI Foundation, a new 501(c)3 non-profit launched by the Cloud Security Alliance at RSA Conference 2026, has taken yet another approach: acquiring and stewarding open specifications rather than publishing guidance documents. In April 2026, CSAI announced that Vanta had contributed the AARM (Autonomous Action Runtime Management) specification—an open system specification for securing AI-driven actions at runtime across dimensions of context, policy, intent, and behavior [6][7]. Herman Errico, AARM's founder, continues to lead development as working group chair. CSAI simultaneously became an authorized CVE Numbering Authority through MITRE, establishing the infrastructure for tracking agentic AI vulnerabilities as a distinct class.

China's response has proceeded through its established technical standardization machinery. TC260, the national cybersecurity standards body, issued a draft practice guide in April 2026 setting out security requirements for deploying and using open-source OpenClaw-class AI agents. The TC260 analysis throughout this note draws on MLex reporting [8]; the primary document is accessible through MLex subscription. The guidance covers the full agent lifecycle—installation, configuration, operational use, and removal—as well as cloud security, supply-chain controls, and organizational governance, including formal requirements for managing "shadow agents" deployed by employees without organizational approval. This guidance follows a series of national standards (GB/T 45654-2025, GB/T 45674-2025, and GB/T 45652-2025) that took effect in November 2025 and established baseline security requirements for generative AI services, data annotation security, and training data governance respectively.

Security Analysis

The Agent Identity Problem Is Structural

Every framework under analysis treats agent identity as a core unsolved challenge in agentic AI security. Current enterprise identity infrastructure—built around human users authenticated via credentials, certificates, and MFA—does not translate directly to AI agents that may spawn dynamically, operate in multi-agent chains, carry delegated authority across system boundaries, and act at machine speed. CoSAI's Workstream 4 explicitly treats AI agents as non-human identities requiring full IAM lifecycle management: access reviews, credential rotation, and privilege certification [4]. NIST's initiative frames the problem as one of "reliably distinguishing AI agents from human users within enterprise systems" and ensuring agents can only act within explicitly delegated scopes of authority [2]. The FIDO Alliance, recognizing the authentication dimension, announced in late April 2026 that it would develop interoperable standards for trusted AI agent interactions, drawing on contributions from Google (the Agent Payments Protocol) and Mastercard (the Verifiable Intent framework) [9]. The convergence of these distinct efforts around the same problem is itself informative: agent identity is not a niche concern but a foundational gap in the current enterprise security stack.

Authorization Drift and Runtime Enforcement

Beyond initial authentication, the harder problem is enforcing bounded authorization throughout an agent's operational lifetime. Agents that begin with narrow permissions can accumulate expanded access through successive tool calls, memory state, and multi-hop delegations in ways that are not visible to conventional access management tooling. AARM addresses this directly: its architecture includes a context accumulator, policy engine, approval service, deferral service, receipt generator, and telemetry exporter, with optional extensions for semantic distance tracking and least-privilege enforcement [7][13]. The framework starts with the action itself—asking how an action should be evaluated against context, intent, policy, and behavior—rather than relying solely on the identity of the requesting actor. This runtime-centric approach fills a gap that identity-focused frameworks leave open: an agent can hold legitimate credentials and still execute unauthorized actions if there is no runtime policy layer capable of evaluating the action in context.

The UC Berkeley Center for Long-Term Cybersecurity published a complementary perspective in February 2026, releasing an Agentic AI Risk Management Standards Profile that extends the NIST AI Risk Management Framework specifically for agentic systems [10]. The profile targets vulnerabilities that emerge from system configuration, tool access, and real-world interaction, including unintended goal pursuit, unauthorized privilege escalation, and the risk of agent self-replication. These failure modes

share surface similarity with conventional privilege escalation and automation bugs but differ in scale, speed, and the difficulty of attributing responsibility across agent chains—requiring governance controls that conventional risk management processes have not been designed to detect.

The MCP Attack Surface

The rapid growth of the Model Context Protocol as the de facto agent-to-tool interface has outpaced security governance. With MCP implemented across more than 10,000 active public servers and over 97 million monthly SDK downloads as of late 2025 [1], the protocol has become a significant attack surface, likely largely unaudited given the pace of adoption. CoSAI's MCP security taxonomy [5] represents one of the first systematic public attempts to categorize the threats this surface introduces: tool poisoning, prompt injection through tool responses, unauthorized data exfiltration via MCP-connected services, and trust confusion in multi-agent MCP chains where one agent's tool output becomes another's input without explicit authorization boundaries. The speed of MCP adoption reflects a broader pattern in enterprise AI: deployment velocity is consistently outrunning security assurance, and organizations that have integrated MCP-connected agents may not have conducted any adversarial assessment of those integrations.

The Geopolitical Fragmentation Risk

The parallel development of AI agent standards across US-led coalitions and China's national standardization bodies introduces a compliance burden that falls specifically on organizations operating in both jurisdictions. China's 15th Five-Year Plan, approved at the National People's Congress in March 2026, places explicit weight on technology innovation, supply chain resilience, and security governance [14], and TC260's OpenClaw guidance signals that China intends to regulate agentic AI deployments through binding technical standards rather than principles-based frameworks [8]. This approach differs structurally from the voluntary, coalition-based model dominant in the United States. Organizations operating in both jurisdictions will face requirements with different ontologies: China's guidance specifies procedural controls at the level of employee handbooks and asset registration forms, while CoSAI's principles focus on architectural governance and shared accountability frameworks. Neither framework is wrong in isolation, but the combination creates compliance overhead before a global consensus has formed. China's broader strategy of leveraging ISO and IEC to advance national standards positions internationally suggests this divergence will not resolve quickly [11].

Framework	Governing Body	Approach	Maturity	Primary Focus
NIST AI Agent Standards Initiative	NIST CAISI (US government)	Standards coordination + research	Early-stage, RFI phase	Interoperability, identity, international standards leadership
CoSAI	OASIS Open (industry coalition)	Principles + operational guidance	Published, iterating	Secure-by-design, identity/access, MCP security
AARM	CSAI Foundation (CSA-affiliated nonprofit)	Open specification	Draft specification, active working group	Runtime action enforcement, policy evaluation
UC Berkeley CLTC Profile	Academic	RMF extension	Published, informational	Risk governance for developers and deployers
China TC260	SAC/TC260 (Chinese government)	Technical practice guides + national standards	Draft guidance, binding standards in force for GenAI	Lifecycle controls, shadow agent management, supply chain

Shadow Agents and Organizational Exposure

Both TC260's draft guidance and CoSAI's governance principles surface the problem of shadow agents—AI agents deployed by employees or teams without organizational awareness, approval, or security review. This category of risk is distinct from shadow IT in that agents carry execution authority: a shadow agent with access to enterprise email, calendar, or code repositories can take consequential actions at scale before the organization is aware it exists. TC260 recommends that organizations establish management systems defining prohibited behaviors, approval processes, and usage boundaries, and that asset registration forms document deployment details, responsible persons, and access scopes for each

approved agent [8]. These controls are practical and achievable with existing process infrastructure. The more challenging problem is detection: without an agent inventory, organizations cannot audit what agents exist, what permissions they hold, or what actions they have taken.

Recommendations

Immediate Actions

Organizations should establish an AI agent inventory as the foundational governance control. This requires cataloging every agent deployment—including agents embedded in productivity tools, coding assistants, and third-party SaaS platforms—with documented ownership, permission scope, and data access. Without this inventory, downstream governance controls such as access reviews and runtime policy enforcement have no reliable target population. Security teams should also treat MCP-connected endpoints as a new class of high-priority attack surface, auditing existing MCP integrations for prompt injection exposure and reviewing tool response handling in any agentic workflow.

Short-Term Mitigations

Agent deployments initiated in the near term should be governed against CoSAI's Principles for Secure-by-Design Agentic Systems [12] as the most operationally mature published standard currently available. CoSAI's principle of treating AI agents as non-human identities with full IAM lifecycle management—including access reviews, credential rotation, and privilege certification—is implementable with existing IAM tooling adapted for service accounts and machine identities [4]. Organizations developing or deploying agents that handle financial transactions or privileged data should additionally track the FIDO Alliance's Agentic Authentication Technical Working Group [9], whose forthcoming specifications for verifiable user instructions and agent authentication may become a foundation for commercial agent authentication standards in the coming years.

Strategic Considerations

Security and governance teams should monitor the NIST AI Agent Standards Initiative closely, particularly its forthcoming guidelines from the NCCoE on adapting identity and authorization frameworks for agents [3], and participate in public comment opportunities when they arise. NIST's work will likely set the compliance floor for US federal contractors and regulated industries. Separately, multinational organizations with operations in China should assess compliance posture against TC260's OpenClaw

guidance now, as draft standards in China's system frequently transition to enforceable requirements on compressed timelines. Organizations should begin developing a dual-track compliance framework rather than waiting for global harmonization. Finally, as AARM matures within the CSAI Foundation, organizations should evaluate adopting AARM-aligned runtime enforcement tooling as part of their agentic security architecture—the specification's action-centric approach fills a gap that neither identity management nor traditional application security controls are designed to address.

CSA Resource Alignment

Note: This research note is published by the Cloud Security Alliance. The frameworks and programs described in this section are CSA products or CSA-affiliated initiatives. Readers should weigh the following recommendations in that context.

The governance landscape described in this note intersects directly with several CSA frameworks and programs. The MAESTRO threat modeling methodology provides a structured approach for identifying agentic AI risks at the layer level, directly applicable to the MCP attack surface analysis and multi-agent authorization failures discussed above. CSA's AI Controls Matrix (AICM) provides domain-level controls mapped to shared responsibility roles—Model Providers, Application Providers, Orchestrated Service Providers—that align with CoSAI's accountability principle requiring that responsibility be delineated before system development begins. The STAR for AI program, including the Catastrophic Risk Annex announced in April 2026, provides a third-party assurance mechanism for evaluating AI deployments against published security controls, enabling organizations to assess vendor compliance systematically rather than relying on self-attestation. CSA holds a seat on CoSAI's Technical Steering Committee, enabling direct contribution to CoSAI's technical workstreams and alignment between CoSAI's principles and AICM controls. The CSAI Foundation's AARM specification, stewarded within the CSA ecosystem, provides a concrete path from governance principles to implementable runtime enforcement through its defined architecture of context evaluation, policy enforcement, and behavioral telemetry. Organizations seeking to operationalize the guidance from NIST, CoSAI, and TC260 should use these CSA frameworks as the connecting infrastructure.

References

[1] Anthropic. "Donating the Model Context Protocol and Establishing the Agentic AI Foundation." Anthropic, December 9, 2025.

[2] NIST. "Announcing the 'AI Agent Standards Initiative' for Interoperable and Secure Innovation." NIST, February 17, 2026.

[3] NIST. "AI Agent Standards Initiative." NIST CAISI, 2026.

[4] CoSAI. "Agentic Identity and Access Management." Coalition for Secure AI, March 2026.

[5] CoSAI. "Coalition for Secure AI Releases Extensive Taxonomy for Model Context Protocol Security." Coalition for Secure AI, 2026.

[6] Cloud Security Alliance. "Cloud Security Alliance Launches CSAI Foundation." CSA Press Release, March 23, 2026.

[7] Cloud Security Alliance. "CSAI Foundation Announces Key Milestones to Secure the Agentic Control Plane." CSA Press Release, April 29, 2026.

[8] MLex. "China's Cybersecurity Standards Body Issues Draft Guidance on OpenClaw AI Agent." MLex, April 2026.

[9] FIDO Alliance. "FIDO Alliance to Develop Standards for Trusted AI Agent Interactions." FIDO Alliance, April 28, 2026.

[10] UC Berkeley CLTC. "Introducing the Agentic AI Risk Management Profile: Expert Perspectives on Governance and Best Practices." Center for Long-Term Cybersecurity, February 24, 2026.

[11] Global Taiwan Institute. "Shaping the Digital Order: China's Role in Technology Standards and the Implications for Taiwan." Global Taiwan Institute, February 2025.

[12] CoSAI. "Announcing the CoSAI Principles for Secure-by-Design Agentic Systems." Coalition for Secure AI, 2026.

[13] Cloud Security Alliance. "AARM: Finding a Path to Secure the Agentic Runtime." CSA Blog, April 30, 2026.

[14] Congressional Research Service. "China's 15th Five-Year Plan: S&T and Economic Priorities." Congress.gov, 2026.