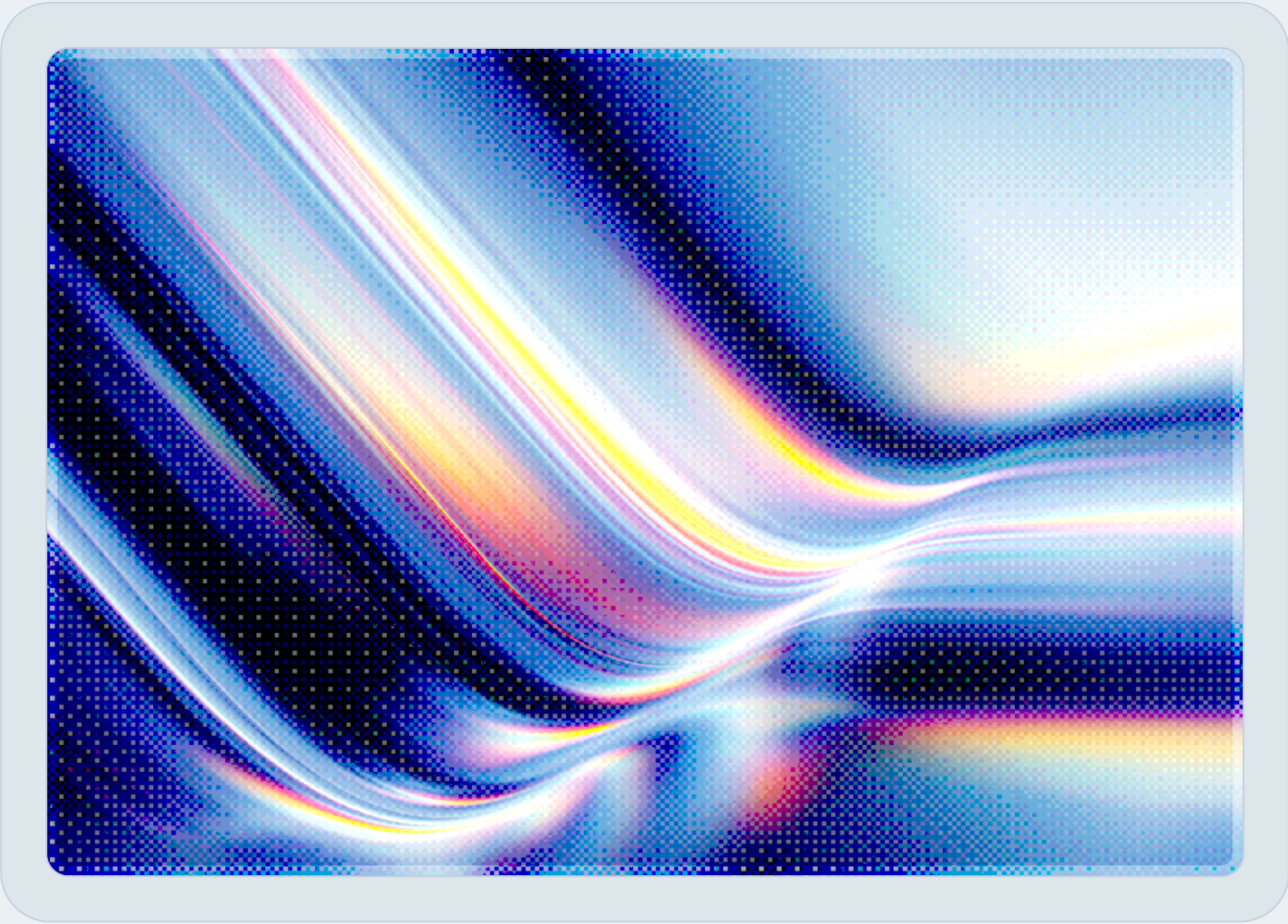
CSAI Foundation | Cloud Security Alliance

Agentic Identity: Emerging Standards and Security Guidance

2026-09-16

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

- Autonomous AI agents have created a new identity category that legacy non-human identity models were not built to handle, because agents are ephemeral, delegate authority across trust boundaries, and act without a human present at the moment of decision.
- Standards bodies acted within months during 2026: NIST's Center for AI Standards and Innovation (CAISI) opened its AI Agent Standards Initiative in February, and the National Cybersecurity Center of Excellence followed with a concept paper examining how OAuth 2.0, OpenID Connect, SPIFFE/SPIRE, and related standards could apply to agent identity and multi-hop delegation [1][2].
- Protocol-level identity is converging on a few patterns: OAuth 2.1 resource-server semantics for the Model Context Protocol, signed Agent Cards for the Agent2Agent protocol, and workload-attested SPIFFE identifiers for infrastructure-level trust [3][4][5].
- Major identity vendors shipped production agent-identity features in 2026, including Microsoft Entra Agent ID and Agent 365, and Okta's and Auth0's MCP-focused authorization layers, signaling that enterprises can begin adopting agent identity governance now rather than waiting for finalized standards [6][7][17].
- Regulators are moving in parallel: Singapore's Cyber Security Agency finalized an addendum to its AI security guidelines in June 2026 that recommends, as voluntary controls, a trusted agent registry, verifiable-credential-based authentication, and human verification for high-impact agent actions [8].
- Organizations should treat agent identity as a distinct discipline from human and traditional service-account identity, and should begin inventorying, registering, and scoping agent credentials well before any single standard is finalized.

Background

The rapid deployment of autonomous and semi-autonomous AI agents across enterprise environments appears to have exposed a structural gap in identity and access management — one that, as the developments below suggest, traditional frameworks were not designed to close. For roughly two decades, non-human identity meant service accounts, API keys, and later workload certificates issued to relatively static, long-lived processes. AI agents break that model in several ways at once. They are

frequently spun up and torn down in minutes rather than months, they reason about which tools and data to access rather than following a fixed code path, and they routinely act on behalf of a human user while also delegating subtasks to other agents, creating chains of authority that existing protocols such as OAuth 2.0, OpenID Connect, and SAML were not built to express. One widely cited industry estimate holds that non-human identities already outnumber human identities by a wide margin in cloud-native environments, with agentic workloads described as the fastest-growing segment of that population [9].

This is not merely a scaling problem. The security literature on agent-to-agent and agent-to-tool communication has begun cataloguing a distinct threat category: identity spoofing and impersonation, in which an attacker exploits weak or absent authentication to make a malicious actor appear as a legitimate agent, a trusted peer, or an authorized user. The OWASP GenAI Security Project's Top 10 for Agentic Applications, released in December 2025 and updated for 2026 after input from more than one hundred security researchers and practitioners, names insecure inter-agent communication and identity or privilege abuse among its top risk categories, noting that when agent-to-agent messages are unauthenticated, unencrypted, or accepted without an origin check, an attacker can impersonate a trusted agent, replay a delegation message to inherit its trust, or register a fake peer in a discovery service to intercept privileged traffic [10]. Because agents are frequently designed to trust collaborating agents by default in order to function autonomously, this class of attack can escalate across a multi-agent system in ways that a human operator may not immediately detect; the OWASP source documents the underlying mechanism, though the pace of detection in practice is this note's own extrapolation rather than a finding the source states directly.

Standards and regulatory bodies responded to this gap over the course of 2026. NIST's Center for AI Standards and Innovation launched its AI Agent Standards Initiative on February 17, organizing work across three pillars covering industry-led standards development, open-source protocol development, and foundational security and identity research [1]. The National Cybersecurity Center of Excellence followed with a concept paper, "Accelerating the Adoption of Software and AI Agent Identity and Authorization," that examines six candidate standards — OAuth 2.0/2.1, OpenID Connect, SPIFFE/SPIRE, SCIM, and NGAC among them — for their applicability to autonomous agents treated as a distinct class of non-human identity, outlining a demonstration project focused on the multi-hop delegation problem broadly, with the Model Context Protocol among the standards under consideration [2]. Internationally, the Cyber Security Agency of Singapore finalized an addendum to its Guidelines and Companion Guide on Securing AI Systems in June 2026, developed with industry, government, and international partners. That addendum recommends, as one set of voluntary controls organizations may adopt, that they maintain a trusted agent registry, authenticate agents using verifiable credentials paired with short-lived tokens, and obtain human verification before an agent takes any high-impact action that affects external systems, modifies persistent data, or triggers a deployment; the addendum's own text states that "these measures and controls are voluntary, and may not be applicable to all organisations and environments"

[8]. Readers should note that the Cyber Security Agency of Singapore and the Cloud Security Alliance share the same three-letter abbreviation; this research note spells out the Singapore agency's full name throughout to avoid confusion with the Cloud Security Alliance's own work discussed below.

Security Analysis

At the protocol layer, three parallel efforts illustrate how the industry is approaching agent identity, each addressing a different part of the problem. The Model Context Protocol, introduced by Anthropic to standardize how agents connect to external tools and data sources, formalized its authorization model around OAuth 2.1 in its June 2025 specification revision. A protected MCP server acts as an OAuth 2.1 resource server, an MCP client acts as an OAuth 2.1 client acting on behalf of a resource owner, and clients must implement Resource Indicators as defined in RFC 8707 so that a token issued for one MCP server cannot be replayed against another [3]. This is intended to close a specific and previously common failure mode in which a broadly scoped token, once obtained, could be presented to any server willing to accept it.

Google's Agent2Agent protocol, donated to the Linux Foundation in June 2025 and now backed by more than one hundred technology companies, including Microsoft, AWS, Salesforce, SAP, and ServiceNow, as project supporters, takes a complementary approach for agent-to-agent discovery and collaboration [4] [11]. Agents publish machine-readable Agent Cards describing their capabilities and the authentication schemes they accept — the protocol's security-object model enumerates API keys, HTTP authentication, OAuth 2.0, OpenID Connect, and mutual TLS — and the specification defines a card-signing mechanism, Signed Agent Cards, so that a receiving agent can verify a card was actually issued by the domain it claims to represent [17]. This addresses discovery-time impersonation, though independent research published in early 2026 concluded that current interoperability protocols, including both MCP and A2A, still lack native mechanisms for verifying the full provenance of a delegated action once it has been handed off through multiple agents [12].

At the infrastructure layer, SPIFFE and SPIRE — an existing open standard and reference implementation for issuing cryptographically verifiable workload identities without long-lived secrets — are being adapted for agentic workloads. SPIFFE's proponents argue it is well suited to agents precisely because it supports automatic rotation and revocation for identities that may exist for only minutes, and because agents already call other agents, tool servers, and model providers through patterns that resemble the workload-to-workload calls SPIFFE was designed for [5]. In this note's assessment, the approach has a known limitation: SPIFFE/SPIRE attestation depends on the identity issuer having visibility into and control over the underlying infrastructure, which breaks down when an agent must present a portable, third-party-verifiable identity to a system operated by an entirely different organization with no shared

infrastructure trust. Emerging work such as the IETF's draft Agent Authorization Envelope attempts to close that gap by layering W3C Decentralized Identifiers and Verifiable Credentials on top of workload identity, so that an agent's authorization can be evaluated by a relying party that has no direct relationship with the agent's home environment [13].

Commercial identity providers moved from research to production during 2026. Microsoft introduced Entra Agent ID, giving each agent created through Azure AI Foundry, Copilot Studio, or other Microsoft surfaces its own manageable identity integrated with Conditional Access, identity protection, and Microsoft Purview, and positioned it as the identity foundation for the broader Agent 365 control plane [6]. Okta and Auth0 responded with agent-focused extensions to their platforms: Auth0's "Auth for MCP" reached general availability in May 2026 [7], and Okta separately released an open-source MCP server that exposes its own management APIs to agents under least-privilege, per-call access control, issuing tokens only with the permissions an operator explicitly grants [19]. The competitive focus appears to be shifting from basic authentication, which is now broadly available, toward governance — inventorying which agents exist, what they are authorized to do, and how that authorization is reviewed and revoked over time.

Taken together, these developments show an industry converging on a layered model rather than a single winning standard: OAuth 2.1 for tool-access authorization, signed capability declarations for agent-to-agent discovery, workload attestation or decentralized credentials for cross-boundary trust, and vendor identity platforms for lifecycle governance. The practical risk for adopters is fragmentation. An enterprise agent may need to satisfy several of these models simultaneously depending on which tools, partner agents, and cloud platforms it touches, and in this note's assessment, the gaps between the models are where identity spoofing and privilege abuse are more likely to occur than within any single, well-specified layer.

Recommendations

Immediate Actions

Organizations operating AI agents today should begin by inventorying every agent identity currently in use, including those created implicitly by agent frameworks that reuse a developer's or service account's credentials rather than issuing the agent its own identity. Where an agent is found to be operating under a shared or overly broad credential, it should be migrated to a scoped, uniquely attributable identity as a priority, since shared credentials make it impossible to attribute an action to a specific agent instance during an incident review. Any Model Context Protocol server exposed to agents should be configured to

require OAuth 2.1 with PKCE and Resource Indicators per RFC 8707, rather than running with authorization disabled, which the specification permits but which removes the primary control against token replay across servers [3].

Short-Term Mitigations

In the next planning cycle, organizations should establish an internal agent registry that records each agent's identity, owner, permitted tools, and permitted counterpart agents. This mirrors the trusted-registry control that Singapore's addendum on securing agentic AI systems recommends, even though that framework treats the control as voluntary rather than mandatory [8]. Inter-agent communication should be authenticated and integrity-checked rather than trusted by default, addressing the insecure inter-agent communication risk that the OWASP GenAI Security Project has identified as a top-tier concern for agentic applications [10]. Organizations piloting SPIFFE/SPIRE or a commercial agent-identity platform such as Entra Agent ID should map out which of their agents operate entirely within infrastructure they control, where workload attestation is sufficient, and which agents must present identity to external parties, where a portable, verifiable-credential-based approach will be needed instead.

Strategic Considerations

Over the coming year, security and identity teams should track the output of NIST's AI Agent Standards Initiative and the National Cybersecurity Center of Excellence's demonstration project, both of which are expected to produce more concrete technical guidance evaluating OAuth 2.0, OpenID Connect, SPIFFE/SPIRE, and related standards for agent identity [1][2]. Organizations with cross-border operations should also monitor how voluntary frameworks such as Singapore's addendum — with its recommended trusted agent registry and human verification for high-impact actions — are received in other jurisdictions, since similar recommendations, and potentially firmer requirements, are a plausible direction for other national frameworks [8]. Finally, because the current standards landscape is genuinely unsettled, organizations should favor architectural choices — such as short-lived credentials, explicit scoping, and centralized registries — that remain compatible with multiple candidate standards rather than committing early to a single vendor's or protocol's identity model.

CSA Resource Alignment

The findings in this research note connect directly to several existing Cloud Security Alliance frameworks. MAESTRO, the Cloud Security Alliance's threat modeling framework for agentic AI, decomposes an agent system into seven layers spanning foundation models through the broader agent ecosystem, with security and compliance running as a cross-cutting layer. In this note's assessment, identity spoofing and delegation abuse map most directly onto MAESTRO's agent framework and agent ecosystem layers, where inter-agent trust and discovery occur [14]. The Cloud Security Alliance's own published guidance on agentic AI identity and access management describes a purpose-built approach using Decentralized Identifiers, Verifiable Credentials, and Zero Trust principles to manage the autonomy, ephemerality, and delegation patterns unique to multi-agent systems, and should be read as a companion to this note for organizations designing a target-state architecture [15][16]. The Cloud Controls Matrix's identity and access management control domain remains the right starting point for mapping specific controls, such as credential lifecycle management and least-privilege enforcement, onto agent identities, even though the matrix's current control language predates the agentic-specific vocabulary this note uses [18]. Organizations pursuing STAR certification or attestation involving agentic systems should anticipate that, as STAR auditors adapt existing IAM control language to agentic systems, evidence requests may extend to agent inventories and delegation records, not solely to human and traditional service accounts. Finally, this note's emphasis on scoped, short-lived, individually attributable agent credentials is consistent with Zero Trust principles that treat identity, rather than network location, as the primary control plane.

References

[1] NIST. "Announcing the 'AI Agent Standards Initiative' for Interoperable and Secure Innovation." National Institute of Standards and Technology, February 2026.

[2] WorkOS. "Everything You Should Know About NIST's AI Agent Standards Initiative." WorkOS Blog, 2026.

[3] Model Context Protocol. "Authorization." Model Context Protocol Specification, 2026.

[4] Authzed. "Agent-to-Agent (A2A) Communication: A Guide to Google's A2A Protocol." Authzed, 2026.

[5] HashiCorp. "SPIFFE: Securing the Identity of Agentic AI and Non-Human Actors." HashiCorp Blog, 2026.

[6] Microsoft. "Announcing Microsoft Entra Agent ID: Secure and Manage Your AI Agents." Microsoft Entra Blog, 2026.

[7] Okta. "Auth0 Gives Developers the Identity Layer to Securely Ship Agentic Apps." Okta Newsroom, May 2026.

[8] Cyber Security Agency of Singapore. "Securing Agentic AI – An Addendum to the Guidelines and Companion Guide on Securing AI Systems." Cyber Security Agency of Singapore, June 2026.

[9] Obsidian Security. "What Are Non-Human Identities? The Complete Guide to NHI Security." Obsidian Security, 2026.

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

[11] Linux Foundation. "Linux Foundation Launches the Agent2Agent Protocol Project to Enable Secure, Intelligent Communication Between AI Agents." Linux Foundation, June 2025.

[12] arXiv. "Security Threat Modeling for Emerging AI-Agent Protocols: A Comparative Analysis of MCP, A2A, Agora, and ANP." arXiv preprint, 2026.

[13] IETF. "Agent Authorization Envelope (AAE): A Machine-Evaluable Authorization Structure for Autonomous AI Agents." IETF Datatracker, 2026.

[14] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance Blog, February 2025.

[15] Cloud Security Alliance. "Agentic AI Identity and Access Management: A New Approach." Cloud Security Alliance, 2026.

[16] Cloud Security Alliance. "Agentic AI Identity Management Approach." Cloud Security Alliance Blog, March 2025.

[17] Agent2Agent Protocol. "Agent2Agent (A2A) Protocol Specification." A2A Protocol, 2026.

[18] Cloud Security Alliance. "Cloud Controls Matrix." Cloud Security Alliance, 2026.

[19] Okta. "Introducing the Okta Open Source MCP Server." Okta Developer Blog, September 2025.

© 2026 Cloud Security Alliance. AI-assisted Rapid Research.