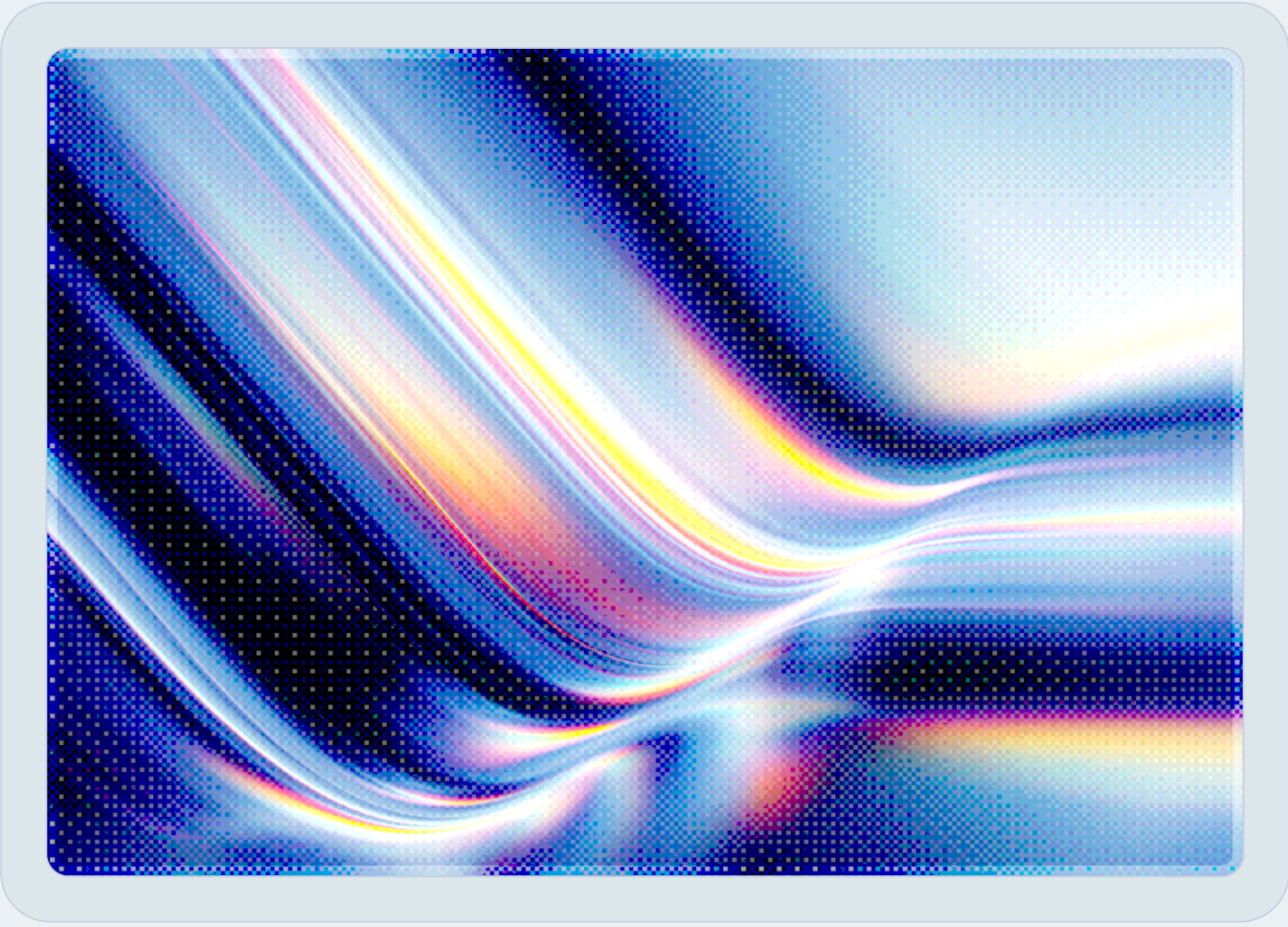
CSAI Foundation | Cloud Security Alliance

Non-Human Identity Management for Agentic AI: Extending IAM Beyond Humans

2026-06-05

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

Agentic AI has moved from experiment to production faster than identity infrastructure has adapted. Survey data from January 2026 found that 73 percent of organizations expect AI agents to become very important or critical to operations within twelve months, yet 68 percent of those same organizations cannot reliably distinguish AI agent activity from human activity in their logs [1][2]. A parallel January 2026 study reported that 92 percent of practitioners are not confident their legacy IAM can manage the new risks introduced by AI agents and non-human identities, and that 78 percent of organizations have no documented policy for creating or removing AI identities [3][4].

Three structural factors account for the distance between deployment speed and security maturity. First, AI agents typically inherit permissions rather than holding their own. Fifty-two percent of organizations represent agents using workload identities, but 43 percent still rely on shared service accounts and 31 percent allow agents to operate under a human user's identity, breaking attribution and least privilege simultaneously [1][2]. Second, traditional identity governance assumes a "joiner-mover-leaver" lifecycle anchored to HR systems that does not apply to identities created dynamically by agent frameworks at runtime [5]. Third, the prompt-injection threat surface converts every inherited credential into a potential leak path: 81 percent of practitioners agree that prompt manipulation could cause agents to reveal sensitive credentials or tokens [2].

Security leaders should treat non-human identity (NHI) for AI agents as an extension of existing IAM discipline rather than a new product category. The control principles are familiar — distinct identity per workload, short-lived credentials, least privilege, attestable lifecycle, and end-to-end auditability — but the implementation primitives are shifting toward cryptographic workload identity (SPIFFE/SPIRE), OAuth 2.0 Token Exchange for delegation, and emerging IETF specifications that bind agent identity to user delegation in a verifiable chain [6][7][8].

Background

Non-human identities now outnumber human identities in enterprise environments by roughly forty-five to one and, in some published estimates, ninety to one, with the ratio growing as automation and AI deployment accelerate [9][10]. CSA's 2024 *State of Non-Human Identity Security* survey of 818 practitioners established a baseline that only 15 percent of organizations were highly confident in their

ability to prevent NHI-based attacks; comparing that baseline with the 2026 follow-up surveys suggests the gap may be widening rather than closing as autonomous agents add a new class of identity on top of legacy service accounts, API keys, OAuth applications, and CI/CD tokens, though the studies used different sponsors, sample sizes, and question designs [9][3]. The principal statistical benchmarks throughout this analysis derive from CSA industry surveys conducted in 2024–2026, each with vendor sponsorship (Astrix Security in 2024, Oasis Security and Aembit in 2026); CSA's research governance applies to all published findings.

What distinguishes an agentic AI identity from the long-standing categories of NHI is not the identity itself but the behavior bound to it. An agent identity is consumed by a process that takes natural-language input, makes routing decisions about which tool to invoke, and chains operations across systems with little or no human checkpoint between the initial prompt and the final action. CSA's *Identity and Access Gaps in the Age of Autonomous AI*, published in March 2026, documents that 67 percent of organizations now use task-automation agents in production, with data retrieval, code generation, security monitoring, and infrastructure operations following close behind [1][2]. Each of these patterns places an autonomous decision-maker between a credential and the resource it accesses.

The Model Context Protocol (MCP), Google's Agent2Agent (A2A) protocol governed by the Linux Foundation, and an IETF Internet-Draft for AI agent authentication (`draft-klrc-aiagent-auth-00`) have all emerged or stabilized in the past year as attempts to standardize how agent identity is asserted, delegated, and verified [6][11][7][25]. MCP mandates OAuth 2.1 with PKCE for human-initiated client flows, while the agent-to-agent and machine-to-machine cases remain works in progress, with client credentials grants and token exchange semantics still under active specification [12][13]. The result is a protocol landscape in which the human-initiated client case has an established specification (OAuth 2.1 with PKCE), while the autonomous agent-to-agent and machine-to-machine cases lack a settled standard.

Security Analysis

The Identity Gray Area

The most cited risk in current practitioner surveys is identity attribution. When 68 percent of organizations cannot clearly distinguish AI agent actions from human actions in logs, the downstream consequences reach across multiple other security capability layers [1]. Incident response cannot determine whether a sensitive action was initiated by a person or an autonomous agent operating under that person's credentials. Forensic timelines cannot reconstruct the causal chain from prompt to action. In most current deployments, detection engineering teams struggle to tune rules for anomalous agent

behavior because agent activity is not reliably distinguishable from human activity in authentication and authorization logs — a direct consequence of the identity conflation described above. Compliance reporting cannot attribute high-impact actions to either a human accountable party or an attested automated process.

This pattern reflects structural choices rather than technical constraints. Organizations that allow agents to operate under human user identities (31 percent) or under shared service accounts (43 percent) have effectively defined the gray area into existence [2]. Even among the 52 percent using workload identity for AI agents, only 36 percent assign a dedicated AI agent identity, and only 22 percent apply their access control framework very consistently across agent deployments [2]. Distinct identity per agent — distinct from any human, distinct from any other agent, and bound to a specific business purpose — is a precondition for all subsequent IAM controls.

Three Structural Gaps

Three properties of agentic systems break the assumptions on which traditional Identity Governance and Administration (IGA) is built. First, **dynamic creation at runtime**. Agent frameworks instantiate sub-agents and tool-calling identities on demand, which means there is no preceding "request" in a ticketing system, no HR record to anchor the identity to, and frequently no human owner explicitly named at creation time. Fifty-one percent of organizations report no clear ownership or accountability for AI identities, and 16 percent do not track when new AI-related identities are created at all [3][14].

Second, **dynamic privilege accumulation**. Agents acquire permissions to complete tasks and tend to retain those permissions long after the task has ended. Seventy-four percent of practitioners agree that AI agents often receive more access than necessary, and 52 percent report that agents inherit access intended for humans or other systems at least sometimes [1][2]. Periodic access reviews — the IGA mechanism designed to catch privilege accumulation — were built for monthly or quarterly cadences applied to a stable population of human users with managers who can answer "does this person still need this access?". The questions appropriate to an agent identity are different: is the identity still in use, what business process depends on it, what is the chain of upstream prompts and downstream tool calls that justify each permission, and is the human owner still accountable [5]?

Third, **delegation opacity**. When an agent calls a tool, which then calls another agent, which then calls a third-party API, the chain of "on behalf of whom" claims is what separates a defensible audit trail from indistinguishable machine traffic. RFC 8693 (OAuth 2.0 Token Exchange) provides the `act` claim precisely to express this chain — the acting party remains distinguishable from the party on whose behalf it is acting — but adoption in agent runtimes remains uneven [8][15]. Where the delegation chain is not preserved cryptographically, every step downstream of the original prompt loses provenance.

Prompt Injection as a Credential Theft Vector

Prompt injection has shifted from a quality issue to a credential exfiltration vector in the past year. Public disclosures in 2026 describe attack classes in which untrusted content — issue bodies, pull request titles, comments, or documents read by a tool-calling agent — instructs the agent to exfiltrate the tokens or API keys present in its execution environment [16]. The "Comment and Control" attack class, disclosed in April 2026 by researcher Aonan Guan and collaborators at Johns Hopkins University, demonstrated a single payload triggering credential theft simultaneously across multiple AI coding agents by abusing GitHub's own commenting surface as the exfiltration channel [24][16]. Wiz documented the "prt-scan" campaign in which more than 500 malicious pull requests targeted GitHub Actions workflows to steal AWS, Azure, and GCP credentials through agent or runner contexts [23][16].

These incidents reframe the identity question. An agent operating with broad inherited permissions is not merely a privilege management problem; it is a *credential proximity* problem. Every credential the agent can read becomes potentially exfiltratable through any untrusted input the agent processes. Eighty-one percent of practitioners agree that prompt manipulation could cause AI agents to reveal sensitive credentials or tokens, and 79 percent agree that AI agents introduce new access pathways that are difficult to monitor [1][2]. The defensive implication is that minimizing the credentials any single agent process holds — through short-lived, narrowly scoped, just-in-time issuance — is more durable than attempting to constrain agent behavior through prompts alone.

The Emerging Standards Landscape

Three standards tracks are converging on the agentic identity problem. The **workload identity** track, anchored by SPIFFE and its SPIRE reference implementation, gives each running workload a cryptographically attested identity (SVID) without static credentials and supports federation with cloud identity providers via OIDC, allowing a workload's SPIFFE ID to be exchanged for short-lived AWS, GCP, or Azure tokens [17][18]. SPIFFE substantially reduces the credential theft attack surface by replacing static long-lived secrets with short-lived, cryptographically attested workload identities. Unlike static API keys, SVIDs expire rapidly and their issuance is tied to a verified workload attestation, making exfiltration significantly less useful to attackers.

The **delegation** track, centered on OAuth 2.0 Token Exchange (RFC 8693), addresses "on whose behalf is this process acting." When an agent operates on behalf of a human user, the token exchange flow combined with the `act` claim preserves the user as subject and identifies the agent as the actor, producing an auditable chain rather than an impersonation [8][15]. An IETF Internet-Draft published in

March 2026 composes WIMSE (Workload Identity in Multi-System Environments), SPIFFE, and OAuth 2.0 into an integrated specification for authenticating and authorizing AI agents in multi-step delegated operations [7][25].

The **authorization** track addresses what an authenticated agent is permitted to do at a given moment. AuthZEN, approved as a Final Specification by the OpenID Foundation in January 2026, standardizes a transport-agnostic API between Policy Enforcement Points and Policy Decision Points, enabling consistent runtime authorization across agent runtimes, gateways, and target systems [26][7]. The MCP authorization specification mandates OAuth 2.1 with PKCE for the human-initiated client case, leaving the autonomous machine-to-machine case to evolving extensions such as client credentials grants and token exchange [12][13]. The A2A protocol publishes an "Agent Card" describing each agent's authentication requirements but explicitly leaves credential provisioning out of scope, which means mutual TLS, signed agent cards, and PKI-backed workload identities remain operator responsibilities [11][19].

Capability needed	Standard / Specification	Status (June 2026)
Attested workload identity (no static secrets)	SPIFFE / SPIRE (CNCF)	Mature; widely adopted
Delegation with auditable actor chain	OAuth 2.0 Token Exchange (RFC 8693), `act` claim	Mature spec; uneven adoption
Composed agent auth across WIMSE + SPIFFE + OAuth	IETF `draft-klrc-aiagent-auth-00`	Internet-Draft (March 2026)
Runtime authorization decisioning	AuthZEN (OpenID Foundation)	Final Specification (Jan 2026)
Human-initiated client to MCP server	MCP Authorization (OAuth 2.1, PKCE)	Required by MCP spec
Agent-to-agent authentication	Google A2A protocol (Linux Foundation)	Adopted; credentialing out of scope

Recommendations

Immediate Actions (next ninety days)

Inventory every place an AI agent currently runs in the environment and record for each: the identity it presents, whether that identity is shared with humans or other agents, the named human owner, the credentials it can read at runtime, and the systems it can write to. The inventory should be authoritative and held outside any single platform vendor, because identity inventory is the dependency on which every later control rests [4][14].

Eliminate the most acute attribution failures by separating agent identities from human identities. The 31 percent of organizations whose agents currently operate under a human user's credentials should be the priority population for this remediation, followed by the 43 percent relying on shared service accounts [2]. Distinct identity per agent is a precondition for least privilege, attribution, and revocation; without it, no downstream control performs as expected.

Run a tabletop exercise simulating prompt-injection-driven credential exfiltration through a production agent. The exercise should validate whether the SOC can detect the exfiltration in logs, whether the affected credentials can be revoked in minutes rather than hours, and whether the blast radius is bounded to the agent's intended scope. Twenty-four percent of organizations currently take more than twenty-four hours to rotate or revoke a credential after potential exposure [3]; that interval is unacceptable for agent-scale operations, where compromised credentials can be exploited within minutes.

Short-Term Mitigations (six to twelve months)

Adopt cryptographic workload identity (SPIFFE/SPIRE or equivalent) for the agent runtime layer, federated through OIDC with the cloud and SaaS resources the agents must reach [17][18]. The objective is to eliminate static long-lived credentials from agent execution contexts and replace them with short-lived attestation-based tokens issued at runtime. This shift substantially reduces the prompt-injection credential proximity risk: even if an agent is manipulated into revealing a token, the token expires in minutes and is scoped to a single workload rather than providing broad access.

Where agents act on behalf of human users, implement OAuth 2.0 Token Exchange with the `act` claim so that audit trails preserve both subject and actor [8][15]. Practitioners should treat the choice between impersonation and delegation in RFC 8693 as a security decision: delegation is the appropriate model for agents because it keeps the agent distinguishable from the user it serves and supports accountable rollback.

Establish an NHI registry that records every AI agent identity's permitted tools, accessible data classifications, lifecycle attestation, and accountable human owner [14]. The registry should be the system of record for access reviews redesigned for non-human identities — usage-based rather than calendar-based, runtime-derived rather than HR-derived, and integrated with the agent platform's telemetry to detect privilege drift in days rather than quarters [5].

Apply zero-standing-privilege principles to the agent population. Just-in-time credentials scoped to a specific task with automatic expiration are an established pattern for human privileged access and translate directly to agents [14]. Where agent platforms support per-task short-lived access, prioritize their adoption; the 20 percent of organizations already rotating credentials continuously or per-use are a reasonable target benchmark [3].

Strategic Considerations (twelve to twenty-four months)

Treat identity as the architectural decision that most constrains future agent flexibility. Whichever system issues, attests, and revokes agent identities will become the substrate on which agent platforms, AI gateways, and policy engines all key off. Organizations that maintain independent identity governance — including for non-human and agent identities — preserve the option to substitute upstream platform components without re-keying the agent population.

Align internal agent identity controls to CSA frameworks rather than vendor-specific configurations. The AI Controls Matrix (AICM) provides the shared-responsibility decomposition needed when one vendor's platform spans multiple control layers; MAESTRO supplies the threat-modeling vocabulary for agent runtimes, tool-calling behavior, and the agent ecosystem layer where identity governance is a cross-cutting concern [20][21]. Framework-anchored policies tend to remain portable across vendor changes; policies expressed entirely in vendor-specific configuration syntax are significantly harder to migrate.

Engage with the standards-track work shaping agent identity. The IETF WIMSE working group, the OpenID Foundation AuthZEN specification, MCP authorization extensions for the autonomous case, and the A2A protocol's evolving security model all influence the practical surface area on which enterprise agent controls will be implemented [7][11][19]. Practitioners participating in these forums are better positioned to align internal architecture decisions with the direction of broader adoption.

CSA Resource Alignment

This research connects directly to several active Cloud Security Alliance programs. The **AI Controls Matrix (AICM)** provides the shared responsibility decomposition and control taxonomy that organizations should use to evaluate whether a platform vendor, model provider, or orchestrated service is meeting its NHI obligations; its identity-related domains align closely with the workload identity, delegation, and lifecycle controls described above [20]. The **MAESTRO** agentic AI threat modeling framework treats identity governance as a cross-cutting concern across its seven layers — particularly L3 (Agent Frameworks) where credentials are issued and consumed, L4 (Deployment and Infrastructure) where workload identities are managed, L6 (Security and Compliance) where access policy is enforced, and L7 (Agent Ecosystem) where credentials traverse multi-agent boundaries [21]. The **Cloud Controls Matrix (CCM)** remains the foundational reference for the underlying cloud-layer identity and access controls that AI workloads inherit, and the CSA **Zero Trust** guidance directly informs the just-in-time, attestation-based, continuously authorized model recommended for agent identities.

CSA's published *State of Non-Human Identity and AI Security* survey and the *Identity and Access Gaps in the Age of Autonomous AI* survey are the principal industry benchmarks for assessing organizational maturity in this area [22][2]. The **STAR** program provides a recognized assurance vehicle for vendor attestation against AICM controls when evaluating third-party AI agent or NHI security platforms.

References

[1] Cloud Security Alliance. "More Than Two-Thirds of Organizations Cannot Clearly Distinguish AI Agent from Human Actions as Over-Privileged Access Becomes Widespread." CSA Press Release, March 24, 2026.

[2] Cloud Security Alliance. "State of AI Agent Identity and Access Security." CSA Research (sponsored by Aembit), March 2026.

[3] Cloud Security Alliance. "The State of Non-Human Identity and AI Security." CSA Research (sponsored by Oasis Security), January 2026.

[4] Cloud Security Alliance. "79% of IT Pros Feel Ill-Equipped to Prevent Attacks Via Non-Human Identities." CSA Press Release, January 27, 2026.

[5] Cloud Security Alliance. "Non-Human Identity Governance: Why IGA Falls Short." CSA Blog, February 5, 2026.

[6] Model Context Protocol. "Authorization Specification." MCP Documentation, 2026.

[7] Riptides. "SPIFFE Meets OAuth2: Current Landscape for Secure Workload Identity in the Agentic AI Era." Riptides Blog, 2026.

[8] IETF. "RFC 8693: OAuth 2.0 Token Exchange." Internet Engineering Task Force, January 2020.

[9] Cloud Security Alliance. "The State of Non-Human Identity Security Survey Report." CSA Research (sponsored by Astrix), June 2024.

[10] The Hacker News. "AI Agents: The Next Wave Identity Dark Matter – Powerful, Invisible, and Unmanaged." The Hacker News, March 2026.

[11] A2A Protocol. "Agent2Agent (A2A) Protocol Specification." Linux Foundation, 2026.

[12] Stack Overflow Blog. "Is That Allowed? Authentication and Authorization in Model Context Protocol." Stack Overflow Blog, January 21, 2026.

[13] Aembit. "MCP, OAuth 2.1, PKCE, and the Future of AI Authorization." Aembit Blog, 2026.

[14] Cloud Security Alliance. "The Non-Human Identity Governance Vacuum (Whitepaper)." CSA Labs, 2026.

[15] Auth0. "Auth0 Token Vault: Secure Token Exchange for AI Agents." Auth0 Blog, 2026.

[16] Waxell. "AI Coding Agent Prompt Injection: CI/CD Risk in 2026." Waxell Research, 2026.

[17] SPIFFE. "SPIFFE Concepts." SPIFFE / CNCF Documentation.

[18] Riptides. "SPIFFE Is What AI Agents Need for Identity, The Question Is How to Deliver It." Riptides Blog, 2026.

[19] A2A Protocol Project. "Agent2Agent (A2A) Protocol Repository." Linux Foundation / GitHub, 2026.

[20] Cloud Security Alliance. "AI Controls Matrix (AICM)." Cloud Security Alliance, 2025.

[21] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[22] Cloud Security Alliance. "AI Agent Identity Crisis: Who's Behind That Action?." CSA Blog, April 20, 2026.

[23] Wiz Research. "Six Accounts, One Actor: Inside the prt-scan Supply Chain Campaign." Wiz Blog, April 4, 2026.

[24] Guan, Aonan, Zhengyu Liu, and Gavin Zhong. "Comment and Control: Prompt Injection to Credential Theft in Claude Code, Gemini CLI, and GitHub Copilot Agent." April 2026.

[25] IETF. "AI Agent Authentication and Authorization (draft-klrc-aiagent-auth)." IETF Datatracker, 2026.

[26] OpenID Foundation. "Authorization API 1.0 Final Specification (AuthZEN)." OpenID Foundation, 2026.