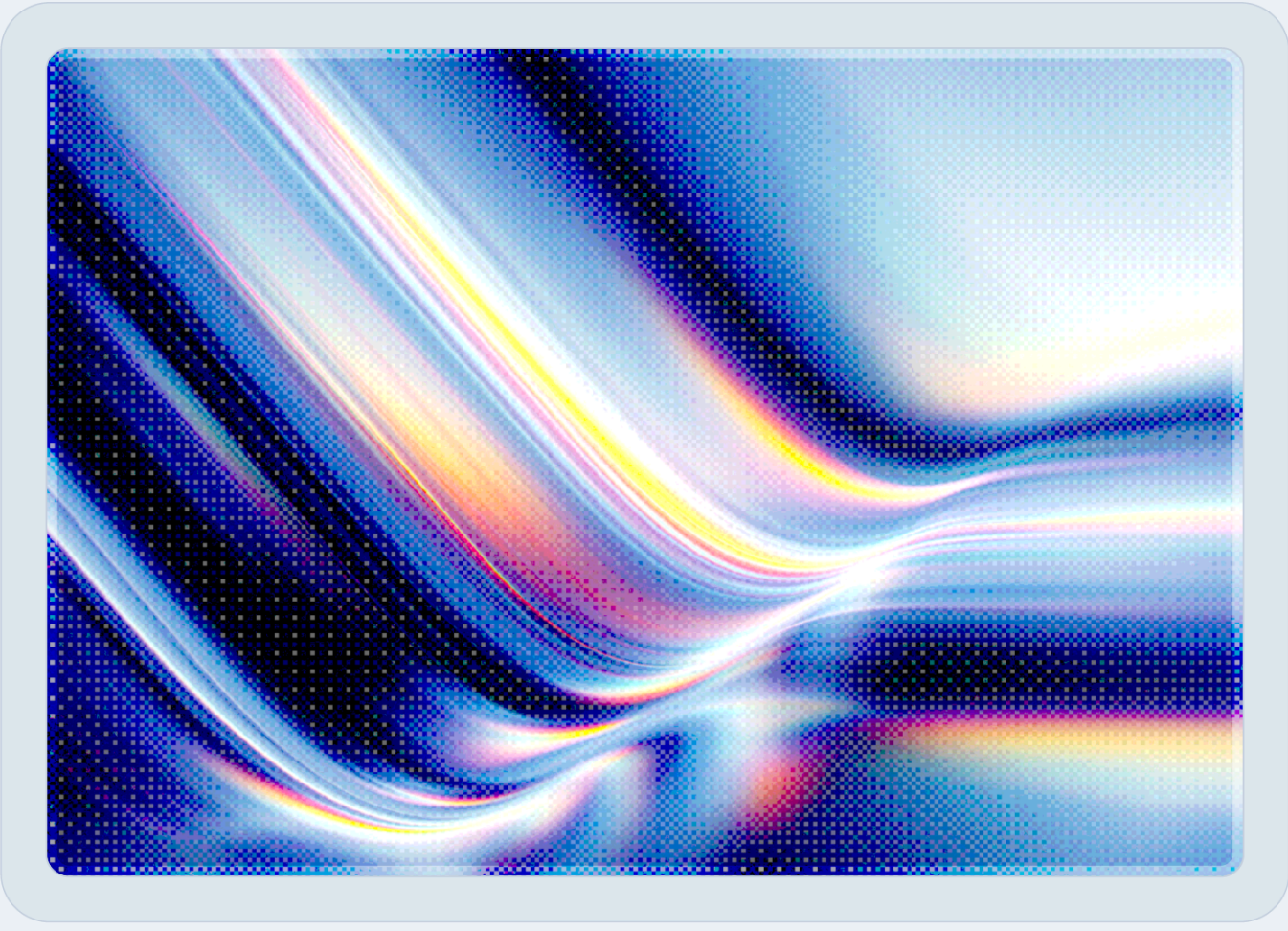
CSAI Foundation | Cloud Security Alliance

AI Agent Identity Sprawl: The Enterprise Authorization Crisis

Why Autonomous Agents Are the Least Audited Identities in the Enterprise

2026-06-20

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Key Takeaways

Enterprise organizations have deployed AI agents broadly—across IT, security, engineering, customer service, and operational workflows—yet the identities these agents hold are managed with less rigor than almost any other class of system in the modern enterprise. According to a January 2026 survey of 228 IT and security professionals conducted by the Cloud Security Alliance and Aembit, 68% of organizations cannot clearly distinguish AI agent actions from human activity, and 74% acknowledge that agents routinely receive more access than their tasks require [1]. These are not theoretical risks: a separate 2026 CSA survey found that 47% of enterprises experienced a confirmed security incident involving an AI agent in the prior twelve months, with 58% of those incidents taking five hours or longer to detect and contain [2].

The core problem is structural. Traditional identity and access management (IAM) was designed around persistent human accounts and durable machine credentials—service accounts, API keys, and certificates managed on a months-or-years lifecycle. AI agents operate differently: they spawn on demand, delegate authority across agent chains, acquire credentials from multiple providers simultaneously, and may be decommissioned without formal offboarding. The result is an emerging class of identity sprawl that most organizations lack the processes and tooling to audit, visualize, or govern comprehensively [1][2].

This research note examines the authorization failure modes most commonly observed in enterprise AI agent deployments, the technical dimensions of the audit gap they create, and the concrete controls organizations must put in place as regulatory and standards bodies accelerate their attention to this domain.

Background

The Identity Gap Takes Shape

For much of 2025, the security conversation around AI agents focused on prompt injection, data exfiltration, and model safety. Identity was treated as a solved problem—an agent is "just another service account." That framing proved insufficient as enterprises began deploying agents at scale.

Non-human identities already vastly outnumber human identities in modern enterprises—industry analyses from major identity vendors consistently estimate between 25 and 50 non-human identities for every human account in cloud-native environments. AI agents add a new and more complex layer: unlike static service accounts, an agent's identity needs may shift mid-task. A research agent aggregating information from Salesforce, a cloud storage bucket, a SaaS API, and a corporate knowledge base may simultaneously hold an OAuth access token, an AWS IAM session credential, an API key, and a bearer token issued by a Model Context Protocol server. These credentials belong to separate trust domains, are visible to different monitoring tools, and expire at different cadences. No single inventory captures all of them together.

MCP, which saw broad adoption across major AI platforms through 2025, emerged as the dominant protocol for connecting AI agents to external tools and data sources, and has accelerated this credential fragmentation [3]. Security researchers have identified significant numbers of internet-exposed MCP servers operating without authentication controls, with the exposure growing as organizations build integrations through both formal and informal channels [3]. As organizations connected enterprise systems to MCP servers—sometimes via individual employees establishing their own OAuth grants—the audit trail for what an agent could access, and under whose authority, became increasingly difficult for security teams to assemble or query comprehensively.

Who Owns Agent Identity?

Accountability for AI agent identity and access is poorly defined in most organizations. The January 2026 CSA and Aembit survey found that no single function dominates ownership: 28% of respondents identified security as the primary owner, 21% cited development and engineering, 19% cited IT, 9% cited IAM teams specifically, and 9% reported no clear owner at all [1]. With ownership distributed across at least four functional groups—and a meaningful fraction of organizations having no designated owner—incident response and routine governance reviews have no authoritative starting point.

This ownership gap has operational consequences. The same survey found that 33% of respondents were unsure how often AI agent credentials were rotated, and 9% reported that credentials were rarely or never rotated [1]. Organizations that cannot identify who owns an agent's identity are unlikely to enforce credential hygiene on it. When 73% of enterprise respondents simultaneously expect AI agents to become critical to their operations within twelve months [1], the gap between adoption velocity and identity governance maturity represents a growing and largely unaudited risk surface.

Security Analysis

Credential Accumulation Without Accountability

A primary authorization risk arising from identity sprawl is credential accumulation—agents acquiring access rights that compound over time and across deployments without any corresponding audit, scope review, or revocation process. This mirrors the service account lifecycle problem enterprises struggled to contain through the 2010s—but now at an accelerated pace, as AI agent deployments have reached organizational scale in months rather than years, and with greater credential complexity due to multi-protocol trust domains.

The January 2026 CSA and Aembit survey documented the specific patterns through which agents accumulate excessive credentials. Forty-three percent of organizations use shared or generic service accounts for AI agents, 31% operate agents under delegated human user identities, and only 36% have established dedicated AI agent identities with their own distinct permission sets [1]. When an agent runs as a shared service account, it inherits the full permissions of that account—often provisioned for a previous purpose—plus any additional grants made for current tasks. The resulting effective permission set reflects historical choices made by multiple teams for different purposes, rather than the minimal permissions the agent's current function requires. Seventy-four percent of enterprise respondents acknowledged that this pattern means agents routinely receive more access than their actual tasks demand [1].

The blast radius of credential compromise scales directly with how broadly a credential is shared. OAuth-based supply chain compromises—in which a single compromised application holds tokens spanning hundreds of downstream customer environments—have emerged as a recurring pattern in enterprise breach investigations, with the scope of access amplification distinguishing these incidents from earlier intrusion types. AI agents that hold similarly broad credentials and act autonomously at machine speed represent an analogous amplification vector, with the additional factor that their capacity for independent action allows compromise to propagate without the friction that human operator involvement would normally introduce.

Attribution Collapse

A second failure mode is attribution collapse: the inability to determine, after the fact, which agent performed which action, under whose authority, and in pursuit of which task. The January 2026 survey found that only 28% of organizations can trace agent actions back to a human sponsor or initiating

context across all environments [1]. In nearly three-quarters of enterprises, agents are operating without a reliably accountable principal attached to their actions.

Attribution failure has implications beyond forensics. Emerging AI governance frameworks, including the EU AI Act's transparency requirements for high-risk systems and NIST's AI Risk Management Framework, increasingly require organizations to demonstrate that system actions can be traced to authorized human decisions. As AI agents perform consequential enterprise operations—scheduling, purchasing, access changes, code deployment—the inability to attach an agent's action to a named human sponsor creates a documented compliance exposure. A 2026 CSA survey of 445 security professionals found that 31% have formally adopted AI agent governance policies, but 50% indicate those policies are only partially documented or inconsistently applied [2].

The technical dimension of attribution failure is not simply a logging problem. Traditional IAM systems can capture OAuth grants and API calls, but they typically cannot record why a tool was invoked, for which agent-level task, and under whose delegated authority. When a multi-agent system—where one orchestrator agent delegates subtasks to specialist agents—generates an API call, the IAM log may show only the credential used, not the delegation chain that authorized its use. The Model Context Protocol's architecture has exacerbated this challenge by enabling tool invocations through a protocol layer that current SIEM and PAM architectures were not designed to parse—most lack native visibility into tool invocations at the MCP layer, a gap security researchers have noted as MCP adoption has accelerated.

Shadow Agent Proliferation

Shadow AI agent deployment—agents introduced into enterprise environments without security team awareness—compounds the identity sprawl problem by creating a population of identities that governance processes cannot even enumerate. The 2026 CSA survey of 445 professionals found that among organizations with more than 100 sanctioned agents, more than half reported an equal or larger number of unsanctioned agents operating in parallel [2]. Shadow agents typically acquire credentials through informal channels: individual employee OAuth grants, personal API keys, or shared development tokens. These credentials exist outside the enterprise's identity governance perimeter by definition.

Shadow agents are not primarily an employee misconduct problem; they commonly emerge because the authorized path to deploy an AI agent is slow or undefined—a structural gap more than a misconduct pattern. The security risk lies not in the intent but in the outcome: agents operating with credentials that no security process has approved, can see, or can revoke.

The Multi-Credential Trust Gap

A more subtle but increasingly important failure mode is what security researchers are calling the multi-protocol authentication gap. An AI agent performing a complex enterprise workflow may simultaneously hold credentials from four or more separate trust domains: an API key for the AI model provider, an OAuth access token for a SaaS application, a cloud provider session credential, and one or more bearer tokens issued by MCP servers or agent orchestration frameworks [3]. Each of these credentials is visible to different monitoring tools—secrets managers, OAuth providers, and cloud IAM consoles each have partial visibility—but no single system provides an integrated view of what the agent can access, how each credential was issued, and when each will expire or be revoked.

This fragmentation means that a security team responding to an incident cannot quickly establish the effective permissions of an agent from any single console. Revoking an agent's access requires coordinated action across multiple systems, and missing any one credential leaves the agent partially operational. In the January 2026 CSA survey, 49% of respondents identified disabling identity or revoking tokens as their primary containment action for AI agent incidents, but the capability to do so comprehensively and rapidly requires an agent identity inventory that most organizations do not have [1].

Recommendations

Immediate Actions

Organizations should begin by establishing an authoritative inventory of all AI agents currently operating in their environment, including agents deployed by business teams outside IT oversight. This inventory should capture, at minimum, the agent's purpose, the credentials it holds, their origin (shared account, dedicated identity, or delegated human identity), the team responsible for it, and the date of last access review. The reality documented by the 2026 CSA surveys is that most organizations cannot produce this inventory from existing tooling; building it requires active discovery across cloud IAM logs, OAuth consent records, API gateway access logs, and developer platform activity.

Parallel to discovery, organizations should implement a credential hygiene baseline for all identified AI agents. Any agent running under a shared service account or delegated human identity should be flagged for remediation—either migrated to a dedicated agent identity with scoped permissions or decommissioned if its business purpose cannot be justified. Credentials that cannot be confirmed as rotated within a defined window should be rotated immediately. The 33% of organizations that are unsure of their agent credential rotation cadence [1] should treat that uncertainty as a finding requiring immediate resolution.

Short-Term Mitigations

Over a three-to-six-month horizon, organizations should establish dedicated identity representations for AI agents as a policy requirement for all new deployments. Each agent identity should be provisioned with the minimum permissions necessary for its declared task, using short-lived credentials wherever the integration supports them. Per-task, time-bounded access is significantly harder to abuse than standing permissions, and the technical underpinnings for it are available today: OAuth 2.0 dynamic client registration and SPIFFE workload certificates are production-ready, while MCP's authorization extensions, still maturing, are sufficient for controlled deployments.

Organizations should also designate a single accountable function for AI agent identity governance. The distributed accountability observed in current surveys—where security, development, IT, and IAM teams each claim partial ownership—is incompatible with coherent policy enforcement. A governance model modeled on Non-Human Identity (NHI) management, which assigns centralized ownership for credential lifecycle decisions while distributing operational use to authorized teams, provides a workable template. Token revocation capabilities should be tested regularly, not assumed: the ability to revoke an agent's access across all credential types in under a defined threshold is an incident response requirement, not a nice-to-have.

Human-in-the-loop checkpoints should be implemented for any agent action that crosses a consequence threshold—financial authorization, access modification, public communication, or changes to production infrastructure. The 2026 CSA and Zenity survey found that 53% of organizations already require human approval for high-risk agent actions [2], but the scope of what qualifies as high-risk should be defined explicitly in policy rather than left to ad-hoc judgment.

Strategic Considerations

Enterprise-Managed Authorization (EMA) for the Model Context Protocol, launched June 18, 2026, allows enterprise IT administrators to provision third-party MCP integrations through an organizational identity provider—eliminating per-user OAuth consent flows and providing centralized revocation capabilities [4]. Okta became the first supported identity provider at launch, with Anthropic's Claude for Enterprise among the first platforms implementing EMA [4]. Organizations evaluating AI agent platforms should treat enterprise identity provider integration as a selection criterion, not an optional feature.

At a standards level, the landscape is moving toward formal requirements. On February 5, 2026, NIST's National Cybersecurity Center of Excellence published a concept paper, "Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization," proposing a demonstration project to develop practical guidance for enterprise AI agent identity management, with specific standards and technologies to be determined through community input [5]. The AI Agent Standards Initiative

announced by NIST on February 17, 2026 establishes the policy intent behind that demonstration [6]. Organizations that establish disciplined agent identity practices now will be better positioned when these standards mature into compliance requirements.

The deeper strategic challenge is architectural. Current evidence suggests that retrofitting IAM tools designed for static human and machine accounts is unlikely to fully address the ephemerality, delegation patterns, and multi-protocol credential requirements of AI agents—though the adequacy of extended existing tooling remains an evolving question as major identity vendors actively develop capabilities in this space. What the evidence makes clear is that purpose-built capabilities are needed: systems that issue short-lived, task-scoped credentials; maintain audit logs that span delegation chains rather than stopping at the credential boundary; and provide real-time visibility into which agents are active, what they can reach, and who authorized their actions.

CSA Resource Alignment

MAESTRO Threat Modeling Framework

CSA's MAESTRO framework (Multi-Agent Environment, Security, Threat, Risk, and Outcome) explicitly identifies Agent Identity Attack as a threat category targeting the identity and authorization mechanisms of AI agents at the Agent Ecosystem layer (Layer 7) [7]. The authorization failures documented in this note—credential accumulation, attribution collapse, shadow agents, and multi-credential trust gaps—all map directly to MAESTRO threat vectors. Security architects conducting threat modeling exercises on AI agent deployments should use MAESTRO's seven-layer decomposition to systematically identify authorization exposure at each layer, from the foundation model through the agent ecosystem.

AICM: AI Controls Matrix

CSA's AI Controls Matrix (AICM) provides a comprehensive control framework applicable to AI agent identity governance. The AICM's identity management, access control, and governance domains contain controls directly relevant to the issues identified in this note, including requirements for dedicated identity assignment, credential lifecycle management, and accountability mapping for AI systems. Organizations implementing AICM controls should specifically evaluate coverage of non-human identity management practices and extend their control mapping to encompass AI agent-specific credential types including OAuth tokens, MCP bearer tokens, and cloud IAM session credentials.

Agentic AI IAM: A New Approach

CSA's 2025 publication "Agentic AI Identity and Access Management: A New Approach" provides architectural guidance for organizations building purpose-built agent identity programs [8]. The publication addresses the use of Decentralized Identifiers (DIDs) and Verifiable Credentials (VCs) as technical mechanisms for establishing agent identity in multi-agent systems, and provides a blueprint for applying Zero Trust principles to agent credential flows. Organizations that have moved past the inventory and baseline hygiene stages described in this note should engage this publication for deeper architectural guidance.

Zero Trust and the Agentic Trust Framework

CSA's Agentic Trust Framework, published in February 2026, extends Zero Trust principles to AI agent governance, providing a governance model for bounded autonomy—autonomy that is tiered, conditional, and tied to verifiable delegation [9]. The framework's emphasis on real-time policy enforcement and context-driven access decisions is directly relevant to the multi-protocol trust gap described in this note. AI agents that can acquire and exercise credentials across multiple trust domains should be governed by policies that treat each action as a fresh authorization decision rather than assuming that a prior credential grant covers all subsequent behavior.

CSA IAM Working Group

The CSA Identity and Access Management Working Group's publications on machine identity, shadow access, and cloud IAM provide foundational context for AI agent identity governance. The Working Group's 2023 paper "Defining Shadow Access" established the conceptual framework for understanding how unintended access pathways arise in cloud environments from automation and DevOps velocity—a dynamic that AI agent adoption is now reproducing at greater scale and speed [10].

References

[1] Hillary Baron, Marina Bregkou, Josh Buker, Ryan Gifford. "Identity and Access Gaps in the Age of Autonomous AI." Cloud Security Alliance and Aembit, March 2026 (survey fielded January 2026).

[2] Hillary Baron, Marina Bregkou, Josh Buker, Ryan Gifford. "Enterprise AI Security Starts with AI Agents." Cloud Security Alliance and Zenity, 2026.

[3] Aembit. "AI Agent Identity: The Multi-Protocol Authentication Gap." Aembit, 2026.

[4] Anthropic. "Centrally manage authorization for MCP connectors." Anthropic, June 2026.

[5] NIST National Cybersecurity Center of Excellence. "Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization (Concept Paper)." NIST CSRC, February 2026.

[6] NIST. "AI Agent Standards Initiative." National Institute of Standards and Technology, February 2026.

[7] Ken Huang et al. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 2025.

[8] Cloud Security Alliance. "Agentic AI Identity and Access Management: A New Approach." Cloud Security Alliance, 2025.

[9] Cloud Security Alliance. "The Agentic Trust Framework: Zero Trust Governance for AI Agents." Cloud Security Alliance, February 2026.

[10] Sasi Murthy Venkat Raghavan, Steven Schoenfeld et al. "Defining Shadow Access: The Emerging IAM Security Challenge." Cloud Security Alliance IAM Working Group, 2023.

