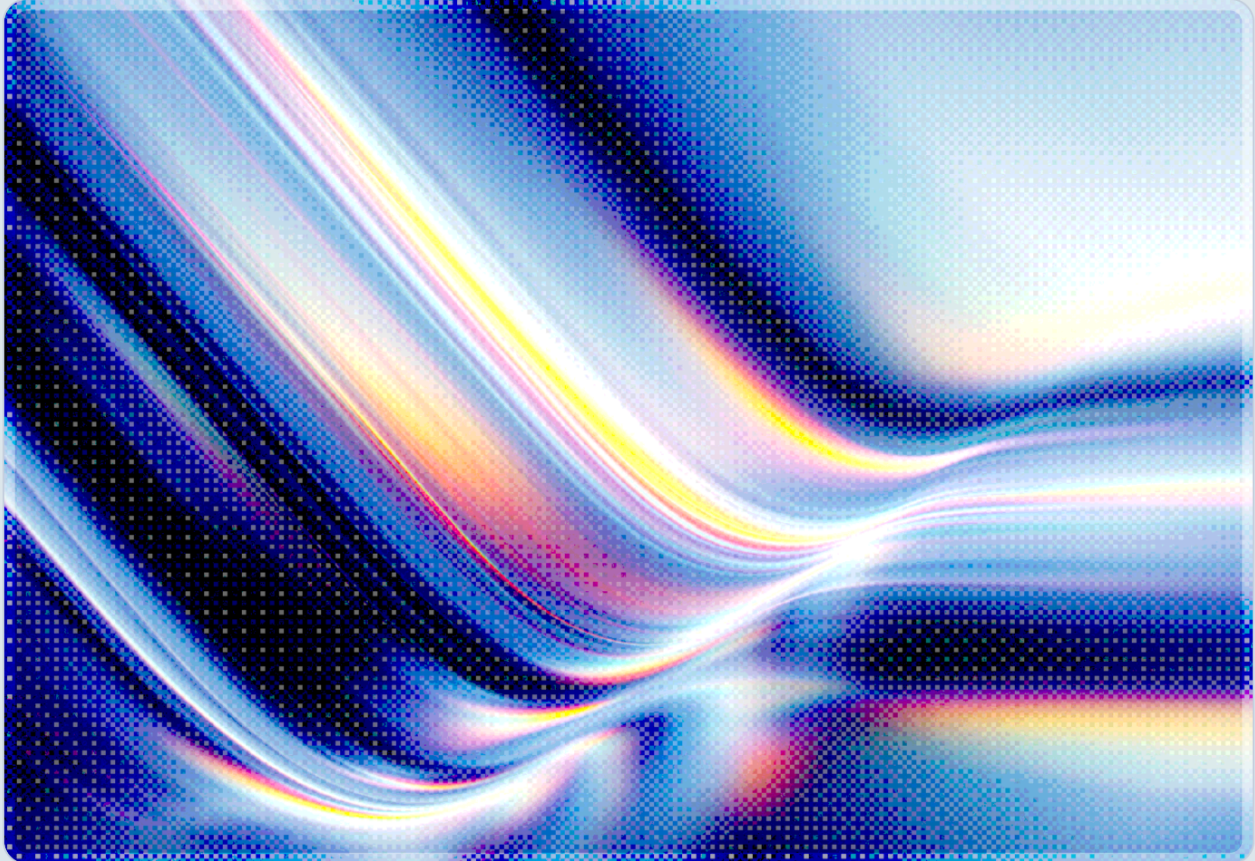


Detecting Shadow AI Agents at the Non-Human Identity Perimeter

2026-05-29

 AI-assisted Rapid Research



© 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

Shadow AI agents – autonomous workflows deployed without security, IT, or governance oversight – have moved from edge case to a near-universal enterprise concern in 2026. A Cloud Security Alliance survey of 418 IT and security professionals conducted in January 2026 and published in April found that 82 percent of organizations have discovered previously unknown AI agents in their environments within the past year, 41 percent of those multiple times, while 65 percent reported AI agent security incidents over the same period [1][2]. Companion research from CSA and Aembit, drawing on 228 practitioners, found that 68 percent of organizations cannot clearly distinguish AI agent activity from human activity in their telemetry, and that 74 percent agree their agents routinely receive more access than they need [3]. The visibility gap is not a perception problem – a separate finding from the Token Security survey reveals that 68 percent of respondents report high confidence in their AI agent visibility while simultaneously reporting that they keep finding agents they did not know about [1].

The architectural reason this gap exists is that AI agents do not slot cleanly into either the human-identity perimeter (workforce IAM, SSO, conditional access) or the conventional machine-identity perimeter (service accounts, workload identities, secret vaulting). Vendor analysis suggests that a single-agent architecture commonly involves nine to twelve distinct NHIs covering the model, the orchestrator, the tool layer, downstream services, and intermediate caches [4][5]. When employees deploy these agents through personal accounts, browser-resident copilots, IDE-embedded coding agents, or Model Context Protocol (MCP) servers, the agents inherit token scopes and data access that the deploying user often does not understand, and they run in ephemeral contexts where conventional endpoint and network telemetry produce only fragmentary traces [6][7]. Detection therefore has to be designed against the non-human identity perimeter rather than retrofitted onto existing workforce or workload controls.

This research note describes the architecture patterns now emerging for shadow-agent detection and governance, mapped to CSA frameworks. The central recommendation is to treat agent discovery as a continuous identity-layer capability – OAuth consent telemetry, NHI inventory, MCP traffic observation, and behavioral baselining – and to wire that capability into the same lifecycle controls (provisioning, scope attenuation, decommissioning) that govern any other privileged identity. Standards and protocol work to underpin this is in flight at NIST, IETF, and the CSAI Foundation, but production controls have to be built on what is available now.

Background

The shadow-agent phenomenon has three intersecting drivers. First, agent creation has been democratized: low-code and no-code agent builders embedded in SaaS suites, browser extensions that wrap any LLM in an automation harness, IDE assistants that scaffold their own tool chains, and MCP servers that expose any backend system as an agent-callable tool have all moved into general availability over the past 12 months [6][8]. Second, the underlying model economy makes personal-account usage cheap and frictionless – a CSA blog analysis published 26 May 2026 cites 47 percent of enterprise AI usage as occurring through personal accounts that sit outside SSO and audit controls [6]. Third, the supporting non-human identity surface has expanded faster than NHI governance has matured: agents acquire OAuth scopes, API keys, personal access tokens, and service principals at the speed of natural-language prompting, and inherit whatever the deploying human can already do [4][9].

Independent surveys converge on the size of the resulting blind spot. CSA's April 2026 report with Token Security ("Autonomous but Not Controlled") found that 82 percent of organizations have discovered unknown AI agents in the past year, 51 percent of those discoveries in internal automation or scripting environments, 47 percent in LLM platforms (including custom tools, assistants, and plugins), 40 percent in SaaS tools with built-in automation, and 40 percent in developer-created workflows [1][2]. A second CSA study with Zenity, "More Than Half of Organizations Experience AI Agent Scope Violations," reported in April 2026 that 53 percent of organizations have had AI agents exceed their intended permissions, 47 percent experienced a security incident involving an AI agent in the past year, and only 15 percent had defined ownership for more than three-quarters of their agents [10]. CSA's January 2026 survey with Aembit on identity and access gaps added that only 36 percent of organizations assign dedicated AI-agent identities, while 43 percent rely on shared or generic service accounts and 31 percent run agents under a human user's identity [3].

The cost dimension is also coming into focus. The CSA and Token Security study reports 61 percent data exposure, 43 percent operational disruption, and 35 percent financial losses among organizations that experienced an AI agent incident [1][2]. Against this backdrop, NIST's National Cybersecurity Center of Excellence (NCCoE) released a concept paper on 5 February 2026, "Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization," to begin a NCCoE-lab demonstration project on agent identification, authorization, and access delegation; the public comment window closed on 2 April 2026 with substantive controls expected later in the year [12][13]. The OWASP Agentic Top 10 for 2026 was published in parallel, naming identity and privilege abuse (ASIO3), insecure inter-agent communication (ASIO7), cascading failures (ASIO8), and rogue agents (ASIO10) among its top concerns [14][15].

Security Analysis

The Non-Human Identity Perimeter

Shadow-agent detection is fundamentally an identity problem because the agent is not a node, a workload, or a session in any sense that legacy perimeters were designed to inventory. A single deployed agent typically composes multiple NHIs: an OAuth client used to authenticate the agent to its tool registry, separate API keys for each downstream tool, a service principal or workload identity for the runtime environment, ephemeral tokens that the agent mints during execution, and – increasingly – credentials that the agent issues to subordinate or peer agents during multi-step reasoning [4][9]. The result is that the "thing" a security team needs to govern is not the agent process but the cluster of identities the agent uses, the consents that bind those identities to data sources, and the runtime call graph that links them.

This is why CSA's research consistently finds that the existing perimeter is mis-targeted. Sixty-eight percent of organizations cannot clearly distinguish AI-agent activity from human activity in their telemetry because the agent is acting under a human's session, a shared service account, or a workload identity that pre-dates the agent's deployment [3]. Only 22 percent of organizations apply their access-control frameworks very consistently to AI agents, and only 25 percent determine agent access through predefined rules – the rest derive it from the requesting human's permissions or the agent's own self-asserted scope [3]. Where the agent inherits a human user's tokens, identity-centric detection has to start by separating agent activity from the human's normal pattern of use.

The OWASP Agentic Top 10 for 2026 names this directly. ASI03 ("Identity and Privilege Abuse") describes the case where agents inherit sessions or reuse secrets across contexts so that privilege escalation and action attribution both fail; its core mitigation is per-agent personas with scoped, time-bound permissions [14][15]. ASI07 addresses inter-agent communication that lacks mutual authentication, schema validation, or encryption, which lets a compromised agent spoof or replay messages across an agent chain [14]. ASI08 ("Cascading Failures") captures the propagation property – a single poisoned memory entry or compromised credential fans out across agents and workflows in ways traditional segmentation models do not contain [14]. Together these failure modes argue for governance built on the identities the agents present, not the network paths they traverse.

Detection Architecture: Five Observation Layers

Practical shadow-agent detection in 2026 is built across five layers, each of which catches a different category of unsanctioned deployment. The first layer is OAuth and consent telemetry. Vendor solutions launched in the first half of 2026 – most visibly Okta's Agent Discovery within its Identity Security Posture Management product, announced 12 February 2026 – detect agents by observing OAuth consent flows initiated against the organization's identity provider and against unsanctioned platforms reached through browser sessions [16]. Because agent builders, plugin marketplaces, and LLM platforms all require OAuth or token grants to access enterprise data, OAuth consent telemetry has emerged as a high-signal first-detection source for new shadow agents, capturing the initial authorization event that most agent deployments require [16][6].

The second layer is the NHI inventory itself. Continuous enumeration of API keys, personal access tokens, service principals, workload identities, and SaaS-issued machine credentials – long a discipline for cloud and DevOps security teams – becomes the substrate for agent attribution once enriched with the upstream consent grant and the downstream resources the credential can reach [4][17]. Where the NHI inventory is owned by a different team than agent governance – and CSA's Aembit survey found 28 percent of organizations placing AI agent identity with security, 21 percent with development, 19 percent with IT, 9 percent with IAM, and 9 percent reporting no owner – coverage gaps appear at exactly the team boundaries [3].

The third layer is MCP and tool-registry observation. The Model Context Protocol has become the dominant integration pattern for agent-to-tool communication, with the ecosystem reportedly hosting more than 13,000 MCP servers and a separate analysis surfacing roughly 8,000 exposed MCP servers without authentication protection [9][18]. Observing MCP and equivalent tool-call traffic at the network or proxy layer lets defenders enumerate tools an agent is actually using rather than tools it was nominally configured for, and surfaces unsanctioned MCP endpoints that employees have stood up on personal infrastructure. CSA analysis of agent-to-agent traffic published 17 April 2026 emphasizes that the runtime call graph – what tools agents invoke, what data they pass, which other agents they activate – is where dangerous combinations of low-risk agents become high-impact attack chains, and is therefore where assurance-grade visibility has to be built [19].

The fourth layer is behavioral baselining of agent execution patterns. Agents make high-frequency API calls that follow structured reasoning loops – the Thought–Action–Observation cycle common in ReAct-style agents is one widely documented example – and the resulting traffic differs from both human and conventional service-account traffic in ways that can be modeled [11][19]. Deviation from declared scope, unusual API call sequences, tool invocations the agent has not historically used, and access of data

classes the agent's declared purpose does not cover are all observable signals; OWASP ASI01 ("Agent Goal Hijack") and ASI10 ("Rogue Agents") both rely on these deviations as primary detection signals [14] [15].

The fifth layer is endpoint-resident agent visibility. A substantial portion of shadow agents now lives inside the developer's IDE, the browser, or the desktop – environments where the agent is ephemeral, runs under the user's identity, and may never produce a server-side authentication event. Endpoint instrumentation that recognizes agent runtimes (LangChain workers, MCP clients, browser-based agent runners, IDE-embedded coding agents) and emits identity-bound execution telemetry closes the gap that the other four layers cannot – but it depends on agent runtimes being recognizable by the endpoint and on the deploying user being subject to managed endpoint controls in the first place [6][7].

Governance Architecture: Bounded Autonomy and Exception-Driven Control

In practice, detection without governance has tended to grow the inventory of known-but-unaddressed agents rather than shrink it – a dynamic that CSA's finding that only 15 percent of organizations have defined ownership for more than three-quarters of their agents [10] makes concrete. CSA's "AI Agent Governance: From Visibility to Control" research, published April 2026, identifies "bounded autonomy" and "exception-driven control" as the two structural patterns that mature governance programs are converging on [2][20]. Bounded autonomy treats autonomy as tiered and conditional – an agent may operate independently within a defined envelope of risk, but escalation outside that envelope triggers human approval, scope attenuation, or termination. Exception-driven control applies governance at decision points rather than as continuous oversight, on the recognition that an agent making millions of API calls per day cannot be reviewed call-by-call.

For both patterns to work, agents must be identifiable as distinct first-class identities. Per-agent identities allow scope attenuation across delegation chains – a parent agent that delegates to a sub-agent passes a narrower token rather than its own credentials – and they allow real-time identity-level enforcement actions when an agent steps outside its envelope. CSA's identity and access survey found that 49 percent of organizations disable identity or revoke tokens as their primary containment action when an AI agent misbehaves, 42 percent terminate the compute environment, and 37 percent block activity through workflow automation [3]. Identity revocation is the fastest of these three controls but only when the agent has a distinct identity to revoke.

The lifecycle dimension is equally important. CSA's April 2026 governance research found that only 21 percent of organizations have formal AI agent decommissioning processes [2]. "Retirement debt" – agents that were never formally onboarded and so are unlikely to be formally retired – accumulates as a class of NHI risk that mirrors orphaned service accounts; in environments where agentic frameworks dynamically create subordinate identities during execution, security teams should anticipate that NHI

accumulation may compound faster than the orphaned service-account problem [2][20]. Lifecycle controls have to span discovery, ownership assignment, scope review, runtime monitoring, and forced retirement, with each step gated by the per-agent identity.

The table below summarizes the relationship between the five detection layers and the governance controls each enables.

Detection Layer	Primary Telemetry	Governance Controls Enabled
OAuth and consent	Consent grants, scopes, client metadata	Discovery, scope attenuation, revocation [6][16]
NHI inventory	API keys, PATs, service principals, workload identities	Ownership assignment, rotation, decommissioning [3][4]
MCP and tool-registry traffic	Tool calls, MCP endpoints, parameter payloads	Tool-level segmentation, unsanctioned-endpoint blocking [9][18][19]
Behavioral baseline	Call sequences, data-access patterns, deviation events	Anomaly alerting, goal-hijack detection, rogue-agent containment [11][14][19]
Endpoint-resident runtime	IDE/browser agent processes, local execution traces	Personal-account governance, ephemeral-runtime attribution [6][7]

Recommendations

Immediate Actions (0–90 Days)

The first 90 days should be spent eliminating the easiest shadow-agent blind spots and establishing an authoritative AI-agent inventory. Organizations that have an Identity Security Posture Management or NHI security capability already in production should turn on OAuth consent discovery against both the corporate identity provider and known unsanctioned platforms, on the recognition that this single source surfaces the largest proportion of net-new agents [6][16]. Where consent-based discovery is not yet available, the equivalent near-term step is a manual inventory exercise driven from existing OAuth logs, SaaS administrative APIs, and developer-platform usage data, targeted at the four environments CSA's research identifies as the most common shadow-agent habitats – internal automation, LLM platforms with custom tools, SaaS-built-in automation, and developer-created workflows [1][2].

Every agent discovered in this initial sweep should be assigned an owner before any other remediation. CSA's survey data shows that 9 percent of organizations report no clear owner for AI-agent identity and access, and a further 33 percent identify owners for only 26–50 percent of agents [3][10]. Ownership assignment is the precondition for every subsequent governance action; agents without owners cannot be onboarded, scoped, monitored against expectation, or retired. Where ownership cannot be determined, the agent should be paused – its identity disabled or tokens revoked – until ownership is assigned, on the same principle the NHI community already applies to orphaned service accounts.

Short-Term Mitigations (3–12 Months)

Within the first year, the focus should shift to operationalizing the five detection layers as a continuous capability rather than a one-time sweep. The MCP and tool-registry layer deserves particular attention because the ecosystem expanded faster than organizations could establish governance for it; an internal scan for unauthenticated MCP endpoints, sanctioned and unsanctioned, plus a policy requiring authentication and authorization on every MCP server reachable from corporate networks, addresses the most-cited specific MCP failure modes [9][18]. Behavioral baselining should be wired into existing SIEM or detection-as-code pipelines, with explicit detections aligned to OWASP Agentic Top 10 categories – goal hijack (ASI01), identity and privilege abuse (ASI03), insecure inter-agent communication (ASI07), and rogue agents (ASI10) – so that the detection engineering work is anchored to a published taxonomy rather than reinvented per organization [14][15].

The scope-attenuation problem deserves a parallel workstream. The CSA and Aembit survey reports that 74 percent of organizations agree agents routinely receive more access than necessary and only 36 percent assign dedicated agent identities [3]. The remediation pattern emerging from CoSAI and the IETF OAuth work is short-lived per-task tokens, scope attenuation through delegation chains so that a sub-agent never holds more authority than its parent, and elimination of standing privilege wherever possible [21]. Organizations should pilot this pattern on a high-risk agent class – typically agents that touch financial systems, customer data, or production infrastructure – and treat the pilot as the reference architecture for the rest of the agent estate.

Finally, decommissioning needs to be a measurable program rather than an exception. With only 21 percent of organizations having formal decommissioning processes, the most leveraged short-term investment is to extend existing NHI lifecycle tooling – token expiration, rotation, deprovisioning workflows – to AI-agent identities, with periodic review cycles that retire agents whose stated purpose has lapsed [2][3]. Organizations that address retirement debt early, while the agent population is still enumerable, avoid the compounding complexity of retiring agents that have already created their own subordinate identities and data access chains.

Strategic Considerations (12+ Months)

The standards landscape will not stabilize before late 2026. NIST's NCCoE concept paper on software and AI agent identity and authorization, released February 2026 and now past its April 2026 public comment window, will produce a practical guide built in NCCoE labs using commercially available technologies; CSA analysis suggests substantive deliverables are unlikely before late 2026 [12][13][22]. The IETF OAuth Working Group has multiple competing drafts – AAuth, AI Agent Auth, Agent2Agent – that consolidate existing standards (OAuth 2.1, SPIFFE, WIMSE, OpenID Connect) but have not yet converged [21]. The CSAI Foundation's Agentic Best Practices program, announced for 2026, will publish identity-first controls for non-human actors, runtime authorization and privilege governance, agent taxonomy and profiling standards, and an open-source tool repository [23][24].

CISOs should build their long-horizon strategy on the assumption that the canonical agent-identity primitive will be per-agent, short-lived, scope-attenuated, and bound to both a human authorizing principal and a documented agent purpose – not because any single standard has been ratified, but because the leading proposals under active development at NIST, IETF, and CoSAI converge on these properties. Building today's controls on top of frameworks that already align to this trajectory – CSA's AICM for the control surface, MAESTRO for threat modeling, NIST's emerging guidance for the federal compliance reference – positions the organization to absorb specific protocol decisions without reworking the underlying governance model. Sector-level information sharing about shadow-agent discovery patterns, common high-risk combination attack chains, and incident postmortems remains undersupplied; participating in CSAI Foundation and ISAC working groups while these standards firm up allows organizations to contribute to and track protocol decisions as they are made, reducing the learning curve when final standards are ratified [23].

CSA Resource Alignment

CSA's AI Controls Matrix (AICM), with 243 control objectives across 18 domains, is the most direct mapping surface for shadow-agent governance because it already articulates the AI-specific extensions of identity management, supply-chain assurance, and lifecycle governance that the shadow-agent problem requires [25]. The MAESTRO threat modeling framework, built on a seven-layer reference architecture spanning foundation models, data operations, agent frameworks, deployment infrastructure, a vertical security and compliance layer, the agent ecosystem, and evaluation and observability, supplies the threat-decomposition lens that lets the five detection layers map cleanly to specific adversary behaviors at specific layers [26][27].

For the identity dimension, CSA's published work on non-human identities and on AI agent identity gaps – including the "Identity and Access Gaps in the Age of Autonomous AI" survey with Aembit, "AI Agent Governance: From Visibility to Control" with Token Security, and the AI agent scope violations study with Zenity – together provide both the empirical baseline and the control taxonomy that the shadow-agent detection program needs to be measured against [1][2][3][10]. The Cloud Controls Matrix (CCM) remains the foundational mapping for cloud-hosted agent infrastructure, and CSA's Zero Trust guidance is directly applicable to the identity-centric enforcement model that bounded-autonomy governance requires.

For organizations beginning shadow-agent programs in 2026, the highest-leverage starting point is to use the AICM as the AI-specific extension of the existing NHI control inventory, MAESTRO to threat-model the agent estate against the seven-layer reference architecture, and CSA's "AI Agent Governance: From Visibility to Control" patterns to structure the lifecycle program. This combination produces an evidence base that is reusable across NIST AI Agent Standards Initiative deliverables when those land, EU AI Act conformity assessment where applicable, and sector regulators that are beginning to extend existing NHI expectations to AI agents.

References

- [1] Cloud Security Alliance. "[New Cloud Security Alliance Survey Reveals 82% of Enterprises Have Unknown AI Agents in Their Environments](#)." CSA Press Release, 21 April 2026.
- [2] Cloud Security Alliance. "[The Shadow AI Agent Problem in Enterprise Environments](#)." CSA Blog, 28 April 2026.
- [3] Cloud Security Alliance and Aembit. "[State of AI Agent Identity and Access Security](#)." CSA Research Report, 2026.
- [4] Entro Security. "[The OWASP Agentic Top 10 2026: What It Means for AI Agents and Non-Human Identities](#)." Entro Security Blog, 2026.
- [5] Astrix Security. "[The OWASP Agentic Top 10 Just Dropped: Here's What You Need to Know](#)." Astrix Security Blog, 2026.
- [6] Cloud Security Alliance. "[Shadow AI Agents: The Insider Threat You're Not Monitoring Yet](#)." CSA Blog, 26 May 2026.
- [7] Vucense. "[Shadow AI Agents: The #1 Enterprise Security Risk in 2026](#)." Vucense, 2026.
- [8] Cloud Security Alliance. "[State of AI Cybersecurity 2026: 92% of Security Professionals Concerned About the Impact of AI Agents](#)." CSA Blog, 27 May 2026.
- [9] Resilient Cyber. "[Identity Is the Agentic AI Problem Nobody Has Solved Yet](#)." Resilient Cyber, 2026.
- [10] Cloud Security Alliance. "[More Than Half of Organizations Experience AI Agent Scope Violations, Cloud Security Alliance Study Finds](#)." CSA Press Release, 16 April 2026.
- [11] Reco AI. "[Top 10 Shadow AI Detection Tools for Enterprise Security in 2026](#)." Reco AI, 2026.
- [12] NIST. "[Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization \(Concept Paper\)](#)." NIST CSRC, 5 February 2026.
- [13] NCCoE. "[New Concept Paper on Identity and Authority of Software Agents](#)." National Cybersecurity Center of Excellence, February 2026.
- [14] NeuralTrust. "[A Deep Dive into the OWASP Top 10 for Agentic Applications 2026](#)." NeuralTrust, 2026.

- [15] Graylog. "[What is the OWASP Top 10 Agentic AI.](#)" Graylog, 2026.
- [16] Okta. "[Okta Secures the Agentic Enterprise with New Tools for Discovering and Mitigating Shadow AI Risks.](#)" Okta Press Release, 12 February 2026.
- [17] Cloud Security Alliance. "[Securing Non-Human Identities in the Age of AI Agents.](#)" CSA RSAC 2025 Presentation, 2025.
- [18] Cloud Security Alliance Labs. "[The AI Agent Governance Gap: What CISOs Need Now.](#)" CSA Lab Space, 3 April 2026.
- [19] Cloud Security Alliance. "[AI Agents Are Talking: Are You Listening?](#)" CSA Blog, 17 April 2026.
- [20] Cloud Security Alliance and Token Security. "[AI Agent Governance: From Visibility to Control.](#)" CSA Survey Report, April 2026.
- [21] Hogan Lovells. "[Shaping the Future of AI Security: NIST Seeking Input on Agent Identity and Authorization.](#)" Hogan Lovells, 2026.
- [22] Cloud Security Alliance Labs. "[Federal Agentic AI Security: NIST's Emerging Standards Initiative.](#)" CSA Lab Space, 2026.
- [23] Cloud Security Alliance. "[Cloud Security Alliance Launches CSAI Foundation With Mission of Securing the Agentic Control Plane.](#)" CSA Press Release, 23 March 2026.
- [24] BusinessWire. "[CSAI Foundation Announces Key Milestones to Secure the Agentic Control Plane.](#)" BusinessWire, 29 April 2026.
- [25] Cloud Security Alliance. "[AI Controls Matrix.](#)" CSA Framework, 2025.
- [26] Cloud Security Alliance. "[Agentic AI Threat Modeling Framework: MAESTRO.](#)" CSA Blog, 6 February 2025.
- [27] Snyk Labs. "[MAESTRO: Layered Threat Modeling for Agentic AI Ecosystems.](#)" Snyk Labs, 2026.