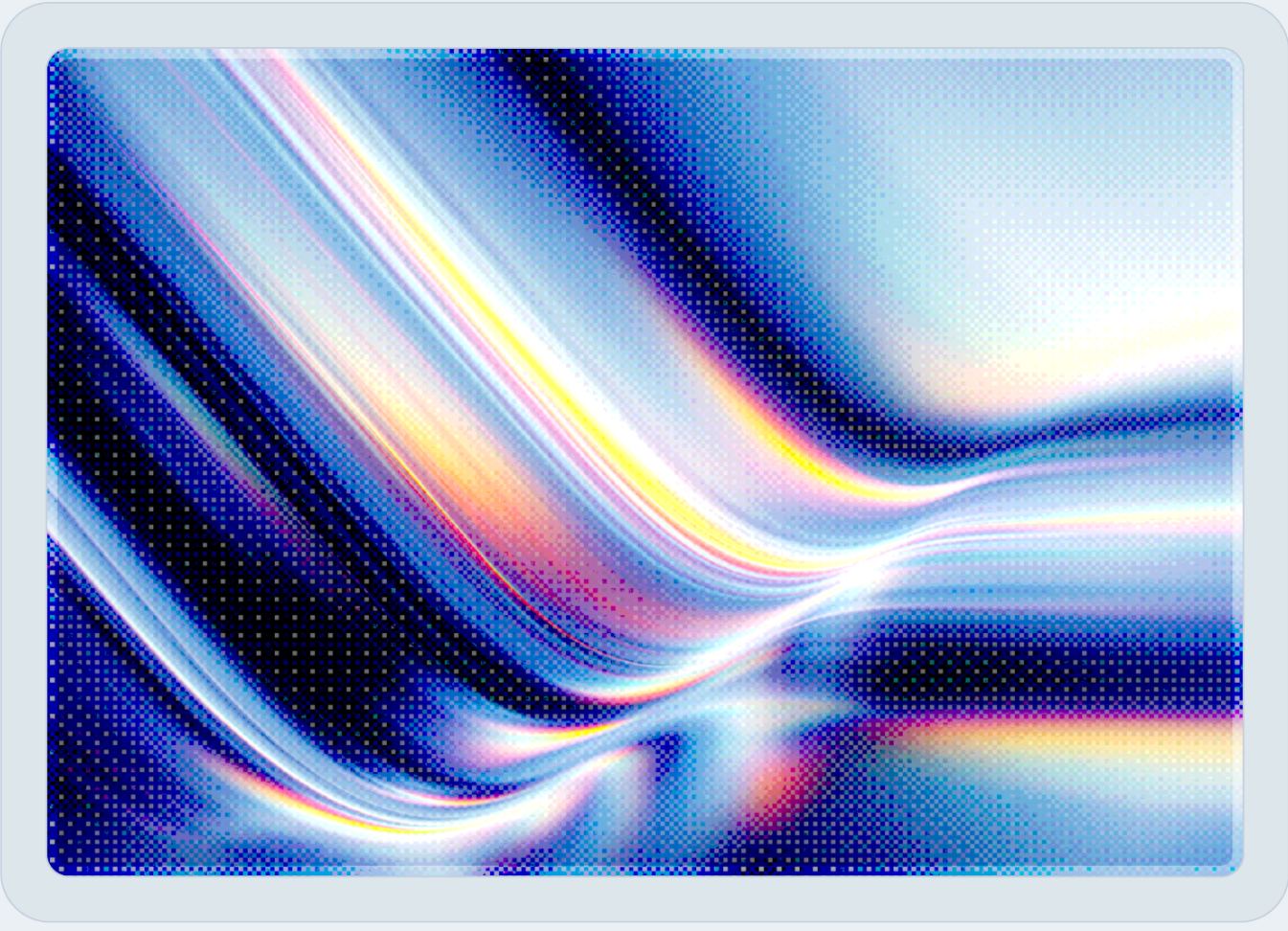
CSAI Foundation | Cloud Security Alliance

Who Did That? Agent Action Attribution as an Audit Obligation

Why regulators, counterparties, and courts increasingly expect enterprises to prove which agent acted, for whom, and under what authority

2026-11-23

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

Most enterprises still treat the ability to say which AI agent performed an action, and on whose behalf, as a security operations concern: useful for forensics, but not a formal obligation. That framing is becoming outdated. Since late 2025, regulators, standards bodies, platform operators, payment networks, and courts have each, in different ways, begun to treat attribution as something an organization may need to *prove* rather than merely good practice. The pressures vary in force, from binding contract terms and existing audit controls to proposed federal guidance, but they point in the same direction. Meanwhile, a CSA survey with Aembit, published in March 2026, found that 68% of organizations cannot clearly distinguish AI agent activity from human activity [1], and 28% reported being able to trace agent actions back to a human sponsor or initiating context across all environments [2][3]. The survey data suggests a substantial gap between emerging expectations and what most organizations can currently produce, and that gap is a compliance and contractual exposure, not only a detection gap.

CSA proposes a four-part evidentiary record for attribution: which agent acted, on whose behalf, under what delegated authority, and what it did. This framing draws most directly on NIST's October 2026 design tenets for agent identity, which warn that if agent activity cannot be linked to the initiating party, organizations face "severe non-repudiation issues" [4]. Regulatory logging duties apply, or are scheduled to apply, to agents operating in regulated systems. U.S. audit controls in HIPAA and NIST SP 800-53 already presume that recorded actions resolve to an identifiable actor [5][6]. Under the EU AI Act, deployers of high-risk systems will need to retain automatically generated logs for at least six months once high-risk obligations take effect, which for Annex III systems is December 2, 2027 following the Digital Omnibus [7][8][9].

Counterparties are moving faster than regulators by writing attribution into contracts and protocols. Amazon's Business Solutions Agreement, updated effective March 4, 2026, requires AI agents to identify themselves as automated systems [10][11], and agentic payment schemes from Visa and Google rely on signed agent identities and signed user mandates [12][13]. Early litigation is already testing attribution questions as well. According to reporting on the decision, the Ninth Circuit's August 2026 ruling vacating a preliminary injunction in *Amazon v. Perplexity* turned largely on whether the user or the agent's vendor was "accessing" Amazon's systems [14]. Against this backdrop, shared service accounts, agents running under human credentials, and agent-writable logs are three common practices that can defeat an attribution claim [1][4], and organizations should begin retiring them now.

Background

Attribution in conventional IT systems rests on a simple assumption: every logged action maps to an authenticated principal, usually a human user or a well-understood service account. Audit frameworks inherited that assumption. NIST SP 800-53 requires that audit records establish the identity of individuals or subjects associated with an event and provides a dedicated non-repudiation control [6]. The HIPAA Security Rule requires covered entities to assign unique user identification and implement audit controls that record activity in systems containing electronic protected health information [5]. These controls were never written with autonomous software in mind. Even so, they do not exempt it. An action taken by an agent is still an action in a regulated system, and the audit trail is still expected to say who performed it.

Agentic AI breaks the assumption in several ways at once. An agent may run under the credentials of the employee who launched it, so its actions are logged as if a human performed them. It may share a service account with dozens of other automations, which collapses distinct actors into one identity. It may delegate work to sub-agents or call tools through protocols such as the Model Context Protocol, adding hops at which the link to the original requester can be lost. The CSA and Aembit survey, fielded in January 2026 across 228 IT and security professionals [1][3], quantified the result: 43% of organizations rely on shared service accounts for agents, 31% allow agents to operate under human user identities, and 74% say agents often receive more access than necessary [1]. CSA's subsequent analysis of these findings named the resulting condition "attribution collapse," meaning an organization cannot determine after the fact which agent performed which action, under whose authority, and in pursuit of which task [2].

What has changed in 2026 is that outside parties are beginning to require the answer. NIST's National Cybersecurity Center of Excellence (NCCoE) opened a project on software and AI agent identity in February 2026, explicitly covering auditing and non-repudiation [15]. On October 27, 2026, NIST published ten proposed design tenets for agent identity management systems. They are not a standard, but they are among the most explicit federal statements to date of what attribution should look like [4]. In parallel, commercial platforms and payment networks have moved from tolerating anonymous automation to demanding that agents declare themselves, and legislators have announced proposals for operator liability for agent conduct [16]. The sections below examine how these pressures translate into concrete obligations.

Security Analysis

What an Attribution Record Must Establish

Drawing on regulatory, technical, and contractual sources, this note frames attribution for an agent action as having four elements. The first is the **agent's own identity**, distinct from both human users and generic service accounts. The second is the **principal**, the human or organization on whose behalf the agent acted. The third is the **authority**, meaning the scope of delegation and the policy or mandate that permitted the action. The fourth is the **action itself**, with enough context to reconstruct it. NIST's tenets supply the most explicit support for this structure. The second tenet requires that agents and the entities initiating their tasks be "cryptographically linked throughout the agent execution," with scopes that attenuate as authority passes to sub-agents [4]. The sixth describes cryptographically signed mandates as providing "an auditable trail for accountability and non-repudiation" [4]. The tenth, "Immutable Evidence," calls for a tamper-evident, append-only record of "the agent's identity, the delegated intent, and the resulting actions," held in logging infrastructure independent of the agent [4].

This four-part structure matters because many existing logs, particularly those of agents running under shared or human credentials, reliably capture only the fourth element [1][2]. CSA's analysis of agentic liability draws the same distinction between output logging, which records what an agent did, and governance logging, which records what policy authorized the action, what delegation the agent operated under, and why [17]. That analysis argues that governance logging is more likely to constitute evidence of reasonable oversight in litigation or a regulatory inquiry, because it shows that the agent operated within authorized bounds rather than simply showing what it did [17]. Commentary on FINRA's 2026 oversight report reaches a parallel conclusion for broker-dealers. Output logs alone may not support reconstruction, and firms may need system-level trails that capture intermediate tool calls [18].

Where the Obligations Come From

The obligations below differ in legal force. Some are binding law, some are contractual terms, and some are proposed or advisory. Taken together, they show attribution moving from optional practice to expected evidence. The table summarizes the principal sources identified in this research, along with each source's status as of the date of this note, so that readers can weigh their legal force independently.

Source	Status (as of November 2026)	What it requires	Attribution implication
EU AI Act Arts. 12, 19, 26 [7][9] [19]	Law; Digital Omnibus amendments, reported finalized in July 2026, postpone Annex III high-risk obligations to December 2, 2027 [8]	High-risk systems must support automatic event logging; providers and deployers retain logs they control for at least six months	Logs must identify events over the system lifecycle in a form usable for oversight and post-market monitoring
HIPAA Security Rule § 164.312 [5]	Law	Unique user identification; audit controls recording system activity	Agents running under a clinician's login produce records that attribute the agent's action to the clinician, undermining audit accuracy
NIST SP 800-53 Rev. 5 (AU-3, AU-10) [6]	Federal control baseline	Audit records identify the subject of each event; non-repudiation for specified actions	Shared agent accounts make subject identification difficult or impossible to satisfy
FINRA 2026 Regulatory Oversight Report [20][18]	Supervisory guidance	Flags agents as difficult to audit; expects monitoring of agent access and data handling	Supervision obligations extend to agent-initiated activity
Amazon Business Solutions Agreement § 19 [10][11]	Contract, effective March 4, 2026	Agents must identify themselves as automated systems and cease access on request	Undeclared agents may breach platform terms
Visa Trusted Agent Protocol;	Industry protocols	Signed agent identity per request; signed user	Merchants and issuers receive verifiable

Source	Status (as of November 2026)	What it requires	Attribution implication
Google AP2 [12] [13]		mandates chained to transactions	evidence of agent and principal
NIST design tenets [4]	Proposed federal guidance	Cryptographic agent–principal linkage; signed mandates; independent append-only evidence	May inform NCCoE reference architectures [15]
AI Agent Accountability Act (proposed) [16]	Announced October 2026; no bill text released	Sponsors state it would expose operators of recklessly harmful agents to CFAA liability	Operators may need records showing who configured and ran the agent

Two points about the table deserve emphasis. First, none of the statutory or regulatory sources was written specifically for agents, yet each applies to agent activity once an agent operates inside a regulated system. An organization that cannot separate agent actions from human actions may struggle to show that its HIPAA audit trail or its SP 800-53 audit records are accurate. Second, the more agent-specific sources are contractual and protocol-based, which means attribution failures may surface first as commercial disputes or chargebacks rather than regulatory findings.

Attribution in Contracts and Disputes

The *Amazon v. Perplexity* litigation is an early example of attribution becoming the central legal question. Amazon alleged that Perplexity's Comet browser agent accessed customer accounts without Amazon's authorization and disguised itself as an ordinary Chrome browser [14]. A district court granted a preliminary injunction in March 2026, and in August 2026 the Ninth Circuit vacated it. According to reporting on the decision, the dispute "largely turned on who was accessing Amazon's servers," and the appeals court found that Amazon was unlikely to establish that Perplexity, rather than the user, was the party accessing its systems, because Comet acts on a user's instructions [14]. The ruling was limited to a preliminary record, and eWeek's analysis notes that future cases may depend on how an agent identifies itself [14]. For enterprises, the practical lesson is that whether an agent is treated as the user's instrument or as an independent third party can determine whose authorization counts. That determination depends on evidence about identity and delegation.

Amazon's contractual response, which requires agents to self-identify under its updated seller agreement [10][11], illustrates how platforms may close this ambiguity through terms of service rather than litigation. Payment networks are taking a similar route. Visa's Trusted Agent Protocol has approved agents sign each request with a registered key, which merchants can verify [12]. Google's Agent Payments Protocol chains signed intent, cart, and payment mandates into what Google describes as a non-repudiable audit trail [13]. These mechanisms put attribution evidence in the hands of counterparties. An enterprise whose agents transact under such schemes is likely to be measured against records the counterparty holds, which raises the cost of weak internal records when a transaction is disputed. CSA's recent analysis of frontier-lab payment wallets observes that consumer protection outcomes under Regulation E and Regulation Z depend on whether a transaction was "unauthorized," a question that strong per-purchase authorization evidence can help settle [21].

Vendor contracts add a further layer of exposure. A 2025 Stanford Law School CodeX guide drawing on TermScout contract data found that 88% of AI vendors impose caps on their own liability and only 17% explicitly commit to complying with all applicable laws [22], findings that CSA's liability analysis also highlights [17]. Where the vendor bears little contractual risk, the deploying enterprise must be able to show for itself that an agent acted within its delegated scope. Attribution records are among the most direct evidence for doing so.

Common Practices That Defeat Attribution

Three common deployment patterns undermine attribution in ways that are difficult to correct after an incident. The first is the **shared service account**. When many agents share one identity, no log entry can distinguish them, and NIST's call for each agent to have its own credential lifecycle cannot be met [4]. The second is **running agents under human credentials**. This may be worse than anonymity, because it creates a record that affirmatively attributes the agent's action to a person. The resulting error can contaminate segregation-of-duties evidence and insider-risk investigations, and it may make a human appear responsible for an action they did not take. The third is **agent-writable logging**. NIST's tenth tenet warns that if activity logs are "accessible or modifiable by the agent itself, a compromised agent could tamper with log files" [4]. An agent that holds the credentials to its own audit store cannot produce trustworthy evidence about itself.

Multi-agent delegation introduces a subtler problem. OAuth 2.0 Token Exchange (RFC 8693) lets a token represent both the subject and the actor through the `act` claim, and nested `act` claims can represent a chain of delegation [23]. However, RFC 8693 itself treats prior actors in such chains as informational only, and an individual Internet-Draft published in March 2026 observes that token exchange does not define how to preserve and validate the full delegation path [23][24]. Observability standards such as the still-maturing OpenTelemetry semantic conventions for generative AI agent spans give useful

telemetry for tracing agent invocations [25]. Telemetry is not tamper-evident evidence, though, and it does not by itself record the authority under which an action occurred. Until standards mature, organizations will likely need to combine token-level delegation, signed intent records, and independent logging to reconstruct multi-hop chains.

Recommendations

The recommendations below are organized by time horizon. They move from discovering where attribution fails today, through near-term changes to identity and logging, to longer-term positioning against emerging standards and contractual expectations. Each step builds on the four-part attribution record described above.

Immediate Actions

Organizations should begin by finding out where attribution fails today. Security and compliance teams should inventory agents that run under shared service accounts or human user credentials, prioritizing agents that touch regulated data, financial transactions, or third-party platforms with agent-identification terms. For each agent, they should confirm that an accountable human owner is recorded. Teams should also test whether, for a sample of recent agent actions, they can answer all four attribution questions from existing records. Where the agent can write to or delete its own logs, those permissions should be removed and log shipping redirected to an independent, append-only store.

Legal and procurement teams should work in parallel on the contractual side. They should review platform terms and payment network rules that now require agent self-identification, beginning with marketplaces and commerce platforms the organization's agents already use. They should also check whether vendor contracts for agent platforms address log access, log retention, and cooperation in incident reconstruction.

Short-Term Mitigations

Over the next one to two quarters, organizations should assign each production agent a distinct workload identity with short-lived credentials, and should express delegation explicitly rather than through credential sharing. Token exchange with `act` claims [23] is a practical starting point. For high-impact actions, the delegation record should be supplemented with a signed statement of user intent, consistent with NIST's mandate tenet [4].

Logging should move from output logging toward governance logging, so that each consequential action records the agent identity, principal, authorizing policy, and tool invoked [17]. Retention should be set to the longest applicable requirement. Organizations with possible EU AI Act high-risk exposure should plan for a six-month minimum [9][19], noting that Annex III obligations now apply from December 2, 2027 [7][8]. Organizations should also add agent-specific scenarios, such as a disputed agent purchase or an agent action under a human login, to incident response and e-discovery playbooks.

Strategic Considerations

Over a six- to twelve-month horizon, organizations should treat NIST's design tenets as a gap-assessment baseline rather than a checklist, since NIST acknowledges that current standards may not yet satisfy every tenet [4]. They should track the NCCoE demonstrations and the IETF's agent-delegation work, and avoid proprietary attribution formats that may not hold up with regulators or counterparties. Contract templates for agent platforms and agent-enabled services should allocate responsibility for producing attribution evidence, not only for liability. Boards and audit committees should be briefed that attribution capability is becoming increasingly important for defending agent-initiated actions, and that operator-liability legislation announced by Senators Hawley and Murphy [16] could raise the stakes further if introduced and enacted.

CSA Resource Alignment

This note builds most directly on CSA's analysis of NIST's agent identity tenets, *NIST's Agent Identity Tenets: A Proposed Federal Reference Point* [26]. That note examines the ten tenets in full. The present note focuses on three of them (agent–principal linkage, signed mandates, and immutable evidence) and argues that external parties are beginning to treat them as evidentiary requirements. Readers implementing the recommendations above should use that note's tenet-by-tenet gap assessment.

The attribution data cited here comes from CSA's *AI Agent Identity Sprawl: The Enterprise Authorization Crisis* [2], which introduced the term "attribution collapse" for this failure mode, and from the underlying survey report *Identity and Access Gaps in the Age of Autonomous AI* [3]. This note extends that work from a security failure mode to a compliance and contractual exposure. On the legal side, *AI Liability Inflection: Enterprise Accountability in the Agentic Era* [17] supplies the "reasonable oversight" framing and the distinction between governance logging and output logging that underpin the recommendations here. CSA's whitepaper *The Non-Human Identity Governance Vacuum* [27] addresses the broader governance of non-human identities in agentic AI, and *IETF's Race to Standardize AI Agent Identity* [28] tracks the standards work relevant to the delegation-chain problem discussed above.

For implementation, CSA's *Agentic AI Identity & Access Management: A New Approach* [29] provides architectural patterns for agent identity and delegation. The AI Controls Matrix (AICM) [30] supplies the control vocabulary for audit. Its Identity and Access Management and Logging and Monitoring domains are the natural anchor points for the controls described above, and organizations pursuing STAR assurance can map their agent attribution evidence to those controls.

References

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

[2] Cloud Security Alliance. "AI Agent Identity Sprawl: The Enterprise Authorization Crisis." CSA AI Safety Initiative, June 2026.

[3] Cloud Security Alliance. "Identity and Access Gaps in the Age of Autonomous AI." CSA, 2026.

[4] Booth, H., Fisher, B., Galluzzo, R., and Clancy, T. "Design Tenets for Agent Identity Management Systems." NIST Cybersecurity Insights, October 27, 2026.

[5] Legal Information Institute. "45 CFR § 164.312 – Technical Safeguards." Cornell Law School, accessed November 2026.

[6] NIST. "SP 800-53 Rev. 5: Security and Privacy Controls for Information Systems and Organizations." NIST, December 2020 (updated).

[7] Gibson Dunn. "EU AI Act Omnibus Agreement – Postponed High-Risk Deadlines and Other Key Changes." Gibson, Dunn & Crutcher, May 2026.

[8] Orrick. "EU AI Act Update: Digital Omnibus Finalizes 8 Compliance Changes." Orrick, Herrington & Sutcliffe, July 2026.

[9] Future of Life Institute. "Article 26: Obligations of Deployers of High-Risk AI Systems." EU Artificial Intelligence Act, accessed November 2026.

[10] Amazon. "Amazon Services Business Solutions Agreement." Amazon Seller Central, accessed November 2026.

[11] Canopy Management. "Amazon's New AI Rules Go Live March 4. Here's What They Mean for Your Business." Canopy Management, 2026.

[12] Visa. "Visa Introduces Trusted Agent Protocol: An Ecosystem-Led Framework for AI Commerce." Visa, October 2025.

[13] Google Cloud. "Announcing Agent Payments Protocol (AP2)." Google Cloud Blog, September 16, 2025.

[14] Ticong, L. "Perplexity's AI Shopping Agent Returns to Amazon." eWeek, August 5, 2026.

[15] NIST National Cybersecurity Center of Excellence. "Accelerating the Adoption of Software and AI Agent Identity and Authorization (Concept Paper)." NCCoE, February 2026.

[16] Office of Sen. Josh Hawley. "Senators Hawley, Murphy Announce Bipartisan AI Agent Accountability Act." U.S. Senate, October 1, 2026.

[17] Cloud Security Alliance. "AI Liability Inflection: Enterprise Accountability in the Agentic Era." CSA AI Safety Initiative, June 2026.

[18] Spitalnick, Fischer, and Melendres. "FINRA's 2026 Oversight Report Signals a Supervisory Reckoning for Autonomous AI." Snell & Wilmer, December 17, 2025.

[19] European Commission. "AI Act Article 19: Automatically Generated Logs." AI Act Service Desk, accessed November 2026.

[20] FINRA. "FINRA Publishes 2026 Regulatory Oversight Report to Empower Member Firm Compliance." FINRA, December 9, 2025.

[21] Cloud Security Alliance. "Agents at the Checkout: Loss Allocation in Frontier-Lab Payment Wallets." CSA AI Safety Initiative, November 2026.

[22] Mack, O. "Navigating AI Vendor Contracts and the Future of Law: A Guide for Legal Tech Innovators." Stanford Law School CodeX, March 21, 2025.

[23] Jones, M., Nadalin, A., Campbell, B., Bradley, J., and Mortimore, C. "OAuth 2.0 Token Exchange (RFC 8693)." IETF, January 2020.

[24] Prasad, Krishnan, Lopez, and Addepalli. "SPICE Actor Chains (draft-mw-spice-actor-chain-01)." IETF Internet-Draft (work in progress; since replaced by draft-mw-oauth-actor-chain), March 16, 2026.

[25] OpenTelemetry. "Semantic Conventions for GenAI Agent and Framework Spans." OpenTelemetry, accessed November 2026.

[26] Cloud Security Alliance. "NIST's Agent Identity Tenets: A Proposed Federal Reference Point." CSA AI Safety Initiative, November 2026.

[27] Cloud Security Alliance. "The Non-Human Identity Governance Vacuum." CSA AI Safety Initiative, 2026.

[28] Cloud Security Alliance. "IETF's Race to Standardize AI Agent Identity." CSA AI Safety Initiative, September 2026.

[29] Cloud Security Alliance. "Agentic AI Identity & Access Management: A New Approach." CSA, 2025.

[30] Cloud Security Alliance. "AI Controls Matrix (AICM)." CSA, 2026.