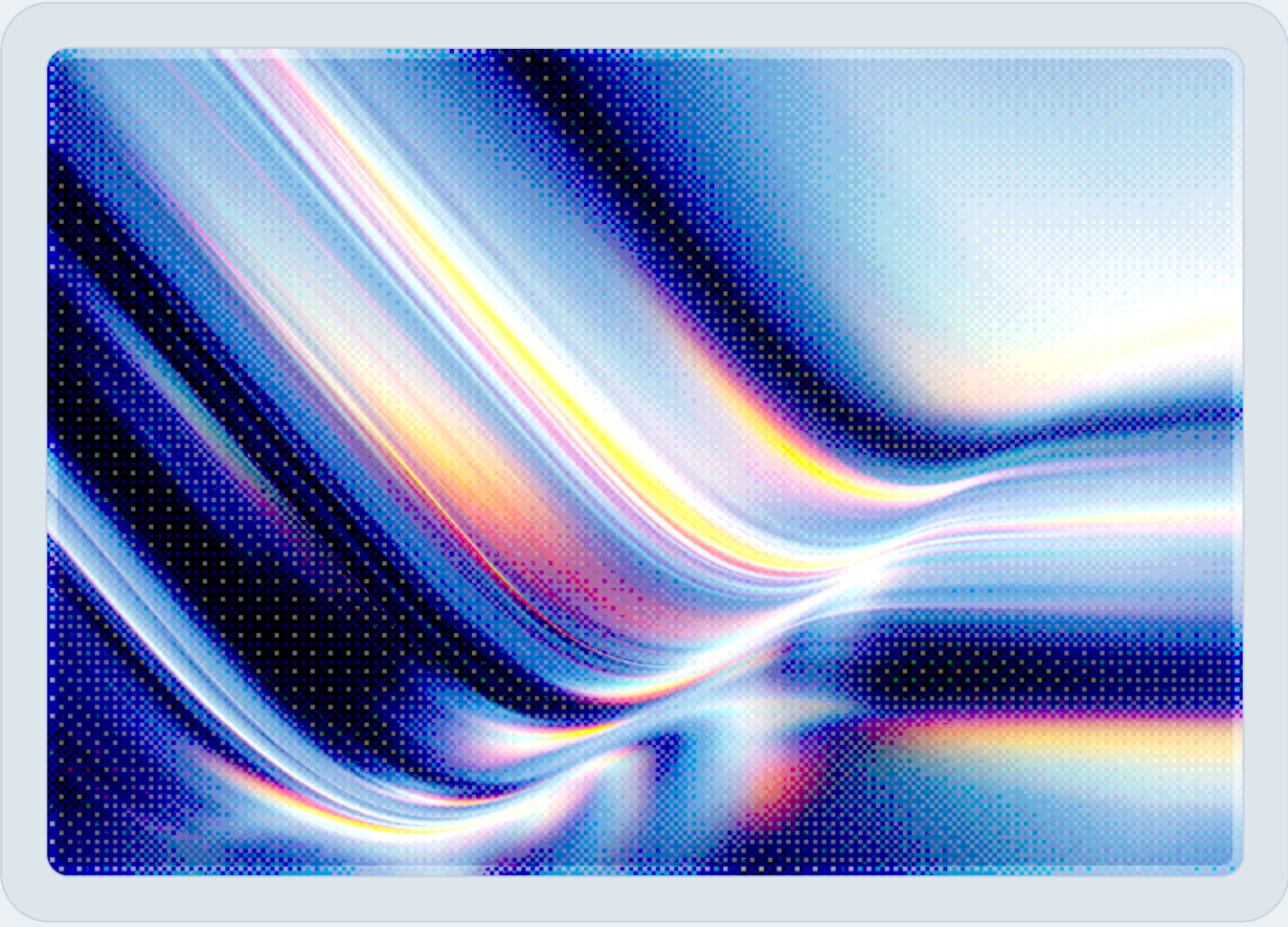
CSAI Foundation | Cloud Security Alliance

NIST's Agent Identity Tenets: A Proposed Federal Reference Point

Security Implications and Guidance for Enterprise Agent Deployments

2026-11-09

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

On October 27, 2026, NIST published ten proposed design tenets for identity, authentication, and authorization of AI agents, authored by Harold Booth, Bill Fisher, and Ryan Galluzzo of NIST and Tom Clancy of MITRE [1]. The tenets are proposed, and they are not a standard or a special publication, but in CSA's assessment they are among the most explicit statements to date of what the U.S. government's identity specialists expect an agent identity management system to do. NIST states that the tenets will guide the National Cybersecurity Center of Excellence (NCCoE) demonstrations of agent identity and authorization, which means they may shape the reference architectures and practice guides the NCCoE produces, and that federal agencies and suppliers may later use as benchmarks [1].

- Agents should be treated as first-class identities with their own credential lifecycle, distinct from both human users and conventional service accounts, and cryptographically linked to whoever initiated their task [1][2].
- Delegated authority should narrow at every hop, agents should hold no standing privileges, and credentials should be short-lived, audience-restricted, and revocable at runtime [1].
- The policy that governs what an agent may do should be architecturally separated from the model that decides what it wants to do; NIST states plainly that "guardrails alone are insufficient" and that deterministic mechanisms should be the primary basis for access decisions [1].
- Human-in-the-loop approval is a necessary but incomplete control, and NIST explicitly warns that poorly designed escalation produces consent fatigue [1][2].
- NIST acknowledges that current standards and technology may not yet satisfy every tenet, so organizations should treat the tenets as a maturity target and gap-assessment baseline rather than a compliance checklist [1].

Background

The tenets follow a series of related federal activities that began in February 2026. On February 5, the NCCoE released a concept paper titled *Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization*, which proposed a project demonstrating how existing identity standards and practices could be applied to software agents, with particular attention to identification,

authorization, auditing, non-repudiation, and controls against prompt injection [3][4]. Less than two weeks later, NIST's Center for AI Standards and Innovation (CAISI) launched a broader AI Agent Standards Initiative organized around industry-led standards, open-source protocol development, and research into agent security and identity [5]. The concept paper's comment period closed on April 2, 2026 [3].

The public response was substantial. According to the NCCoE's published summary, the concept paper drew 614 responses, roughly 70 percent of them from industry and the private sector [6]. The summary records that most respondents favored building on existing identity standards rather than inventing an agent-specific stack, and it identifies recurring themes that closely anticipate the tenets: verifiable non-human identities, stable trust anchors paired with short-lived credentials, attenuation of authority along delegation chains, interest in cryptographically signed statements of user intent, strong support for deterministic policy enforcement, separation of the control plane from the data plane, and tamper-evident audit records [6]. The summary also lays out next steps, including a draft project description, an initial enterprise-focused phase built around securing agents in a DevSecOps software factory, and a later phase addressing consumer-owned agents [6].

NIST then published a two-part blog series that translates those comments into design guidance. The first post, "Back to the Future: Why Agentic AI Needs a Strong Identity Foundation" (August 27, 2026), argues that most agentic identity problems are long-standing identity and access management problems amplified by the speed and scale of agents [2]. It singles out several common practices as hazardous, including handing personal or enterprise credentials to agents, embedding static API keys and bearer tokens in configuration and markdown files, running agents under a user's local account, and relying on frequent human approval prompts [2]. It also names the standards NIST considers foundational, among them OAuth 2.0, SPIFFE, JSON Web Tokens, X.509, Demonstrating Proof-of-Possession (DPoP), Rich Authorization Requests, Transaction Tokens, OpenID AuthZEN, and the Zero Trust architecture described in NIST SP 800-207 [2][7]. The second post, published October 27, introduces the ten tenets themselves [1].

Security Analysis

The Ten Tenets at a Glance

The tenets can be grouped into three clusters: those that establish who an agent is and on whose behalf it acts, those that bound what it may do, and those that make its behavior observable and accountable. The table below summarizes each tenet and the primary risk it addresses, drawing on NIST's descriptions [1].

#	Tenet	Core Requirement	Primary Risk Addressed
1	Distinct Agent Identification	Unique, cryptographically verifiable agent identity, separate from users and service accounts, bound to workload, environment, and approved configuration	Impersonation; loss of traceability
2	Linking, Delegation & Attenuation	Cryptographic link to the initiating person or entity; delegated scope narrows at each hop	Repudiation; unbounded downstream authority
3	Least & Zero Standing Privilege	No persistent rights when idle; task-minimal scope; separation of duties	Lateral movement; privilege accumulation
4	Short-Lived, Dynamic & Revocable Authorization	Just-in-time, time-limited, audience-specific access with continuous runtime evaluation	Token theft and replay; persistent unauthorized access
5	Human-in-the-Loop Triggers	Risk thresholds trigger escalation to the right authority with sufficient context; design for consent fatigue	High-impact autonomous errors
6	Mandates and Constraints	Signed mandates from task originators define intended outcomes and constraints	Unbounded task execution; weak audit trail
7	Separation of Policy from Reasoning	What an agent tries to do is architecturally separated from what it is permitted to do	Prompt injection; privilege escalation
8	Deterministic Access Decisions	Deterministic mechanisms are the final authority; probabilistic signals are inputs only	Manipulation of model-based decisions
9	Continuous Monitoring	Agent-specific metrics, thresholds, and policies for anomaly detection	Undetected compromise;

#	Tenet	Core Requirement	Primary Risk Addressed
			configuration drift
10	Immutable Evidence	Tamper-evident, append-only record of identity, delegated intent, and actions, held independently of the agent	Log tampering; audit failure

Identity Is Being Redefined as a Chain, Not an Account

A central shift in the tenets is that an agent's identity is not a single credential but a verifiable chain linking a workload, its approved configuration, and the human or organization on whose behalf it acts. Tenet 1 binds the agent to a workload, an environment, and an approved configuration, which implies that a modified agent, or one running somewhere unexpected, should no longer be able to assert the same identity [1]. Tenet 2 extends that binding outward to the initiator and then constrains every subsequent delegation so that a sub-agent never holds more authority than its parent [1]. Taken together, these requirements describe something that conventional service accounts were never designed to provide. A service account typically represents an application rather than a principal, persists indefinitely, and carries no record of who asked it to act. The tenets suggest that organizations still provisioning agents through generic service accounts or shared API keys are operating outside the emerging federal expectation, a conclusion consistent with the practices NIST criticized in its August post [2].

The standards landscape partially supports this model today. OAuth 2.0 Token Exchange (RFC 8693) already defines a mechanism for representing delegation and actor chains in tokens, and DPoP (RFC 9449) sender-constrains tokens so that a stolen token cannot simply be replayed by another party [8][9]. The NCCoE comment summary describes respondents converging on a compositional stack rather than a single specification, with workload identity from SPIFFE and the IETF WIMSE effort, delegation through token exchange and transaction tokens, policy evaluation through engines such as NGAC, Cedar, or Open Policy Agent, and decision interfaces through OpenID AuthZEN [6]. What remains unsettled is the cross-organization case. Attenuation is tractable within a single authorization server, but preserving a verifiable delegation chain when an enterprise agent calls a partner's agent, which in turn calls a SaaS tool, depends on standards that are still in draft, such as the IETF Workload Identity in Multi-System Environments (WIMSE) work and related agent authentication proposals [2][13].

Separating Policy from Reasoning Is the Defense Against Confused Deputies

Tenets 7 and 8 carry the most direct implications for how agent platforms are architected. NIST argues that the reasoning engine, which processes untrusted data and is therefore susceptible to prompt injection, must not be the component that decides whether an action is permitted [1]. Probabilistic signals such as model-based risk scores may inform a decision, but the final determination should come from a deterministic mechanism. This position rejects a pattern seen in many current deployments, in which the agent is given broad credentials and a system prompt instructing it to behave, with safety resting on the model's compliance and on output filters; NIST describes the sharing of credentials with agents as one that "has quickly become a common pattern" [2], and CSA's research on agent identity sprawl documents its operational consequences [12].

NIST uses the confused deputy attack to illustrate why individual tenets are insufficient on their own. In that scenario, injected instructions persuade a broadly scoped agent to take actions it would not otherwise take, and NIST states that mitigating it requires tenets 2, 3, 7, and 8 in combination [1]. If authority is attenuated per task, the agent holds no standing privileges, the policy decision point sits outside the model's control, and that decision point evaluates deterministic rules, then a successful injection is confined to actions within the agent's narrowly delegated task scope. Requests outside that scope will be denied by policy, although misuse of in-scope permissions remains possible and calls for continuous monitoring (tenet 9) and human escalation for high-risk actions (tenet 5). Each of these controls is weaker in isolation: least privilege alone still leaves the in-scope action available, and an external policy engine is of little value if the agent's credentials allow it to bypass that engine. Security architects may therefore find it useful to evaluate agent platforms against tenet combinations mapped to specific threats rather than scoring each tenet independently.

Human Oversight Is Reframed as a Risk-Tiered Control

The tenets treat human-in-the-loop approval as a limited control. Tenet 5 calls for organizations to define risk thresholds that trigger escalation, to route escalations to the appropriate authority, and to supply enough context for a meaningful decision [1]. NIST explicitly identifies consent fatigue, in which approvals become so frequent that they are granted by habit rather than judgment, as a design failure to be avoided [1]. The August post goes further, noting that elicitation requests asking users for credentials can themselves enable impersonation, and it points to the Model Context Protocol specification's warning against using elicitation for sensitive information [2]. For practitioners, this framing implies that the number of approval prompts is not a measure of safety. An approval step that fires on every tool call may provide less real oversight than a small number of well-contextualized escalations reserved for irreversible or high-value actions.

Gaps and Open Questions

NIST is candid that existing standards and technology may not yet meet every tenet across all deployment models, and it positions the tenets as input to maturity roadmaps rather than as immediately testable requirements [1]. Several gaps are apparent from the primary sources. Tenet 6's signed mandates have broad support, but the NCCoE summary records no consensus on a standard format or on how to manage the associated privacy trade-offs [6]. Consumer scenarios raise their own tension: NIST notes that consumer transactions may be anonymous or pseudonymous and that mandates should avoid unnecessary disclosure of personal information, which complicates the strong cryptographic linking described in tenet 2 [1]. Tenet 10 assumes an enterprise able to maintain an evidence store independent of the agent, a condition that may not hold for agents operating across organizational boundaries. These are open design problems rather than defects, but they suggest that organizations should expect the tenets to evolve as the NCCoE demonstrations proceed.

Recommendations

Immediate Actions

Organizations deploying agents should begin by using the ten tenets as a gap-assessment instrument against their current agent inventory. The most urgent findings are likely to involve the practices NIST has already identified as hazardous: agents running on shared human credentials, long-lived API keys stored in configuration or prompt files, and agents operating under a user's local account [2]. Each of these should be inventoried and assigned an owner, and any static secret discovered in agent configuration, markdown instruction files, or logs should be rotated. Security teams should also confirm that agent actions are recorded somewhere the agent itself cannot modify, since a compromised agent with write access to its own audit trail can conceal its actions and defeat after-the-fact accountability, which is why tenet 10 requires evidence to be held independently of the agent [1].

Short-Term Mitigations

In the near term (for many organizations, within one to two quarters), organizations should move agent authentication from static secrets to workload identity and short-lived, sender-constrained tokens. In practice, this means issuing agents distinct identities through a workload identity system such as SPIFFE, exchanging those identities for narrowly scoped access tokens at task time using OAuth token exchange, and binding tokens with DPoP or mutual TLS certificate binding so that theft does not confer reuse [2][8][9][10]. Authorization decisions for agent tool calls should be routed through an external policy decision

point rather than embedded in prompts, consistent with tenets 7 and 8 [1]. Human approval workflows should be reviewed against tenet 5: approvals should be reserved for actions crossing defined risk thresholds, and the approval request should show the originating principal, the delegated scope, and the specific action under consideration.

Strategic Considerations

Longer-term planning should allow for the possibility that the tenets, or successors to them, become federal expectations. NIST has said the tenets will guide NCCoE demonstrations [1], and the NCCoE plans a draft project description as its next step [6]. CSA expects that organizations selling to federal agencies, or aligning their programs with NIST guidance, will increasingly see agent identity, delegation, and evidence requirements in procurement questionnaires and assessment frameworks, although no such requirement has been announced. Organizations should also engage with the NCCoE process directly; the August post invites comments at AI-identity@nist.gov, and the forthcoming draft project description will provide a further opportunity to influence scope and architecture [2][6]. Platform selection deserves particular attention. Agent frameworks and gateways that cannot externalize authorization decisions, propagate delegation context, or emit tamper-evident records will be difficult to bring into line with tenets 2, 7, and 10 later, so these capabilities may be worth treating as selection criteria now rather than retrofit requirements.

CSA Resource Alignment

CSA's published work on agent identity addresses many of the same themes as the tenets and offers implementation detail that complements NIST's design-level framing. *Agentic AI Identity & Access Management: A New Approach*, published by CSA in 2025, argues that traditional IAM protocols designed for human users and static applications are poorly suited to autonomous, ephemeral agents and proposes an agent-centric model built on verifiable credentials, decentralized identifiers, fine-grained just-in-time access, and zero trust principles [11]. Its emphasis on dynamic, context-aware authorization and verifiable agent identity maps closely to tenets 1, 3, and 4, and organizations looking for concrete architectural patterns behind NIST's principles will find it a useful companion.

CSA's research note *AI Agent Identity Sprawl: The Enterprise Authorization Crisis* (June 2026) documents the operational conditions the tenets are designed to correct, including credential accumulation across trust domains, attribution collapse when agents act on shared credentials, and the proliferation of unaudited shadow agents [12]. Its findings illustrate the operational conditions that distinct identification, delegation linking, and immutable evidence are intended to address. Similarly, CSA's analysis *IETF's Race*

to Standardize AI Agent Identity (September 2026) tracks the IETF drafts on workload identity, delegation, and agent authentication that will determine how quickly the standards gaps NIST acknowledges can be closed [13]. Read together with the tenets, it helps practitioners distinguish which requirements can be met with ratified specifications today and which depend on work still in progress.

For control mapping and assurance, the CSA AI Controls Matrix (AICM) v1.1 offers a control vocabulary that organizations can use to assess their alignment with the tenets [14]. A formal AICM-to-tenet mapping has not yet been published; in the meantime, its Identity and Access Management domain addresses credential lifecycle, least privilege, and access review; its Logging and Monitoring domain addresses the evidence and monitoring obligations in tenets 9 and 10; and its shared-responsibility model helps allocate tenet obligations among model providers, orchestrated service providers, application providers, and AI customers. CSA suggests that organizations already using AICM for STAR for AI assessments could extend those assessments with tenet-specific questions about delegation attenuation and policy-reasoning separation, which the current tenets make explicit.

References

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

[2] Fisher, B. and Galluzzo, R. "Back to the Future: Why Agentic AI Needs a Strong Identity Foundation." NIST Cybersecurity Insights, August 27, 2026.

[3] NIST. "New Concept Paper on Identity and Authority of Software Agents." NIST News, February 5, 2026.

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

[5] NIST. "Announcing the AI Agent Standards Initiative for Interoperable and Secure Innovation." NIST News, February 17, 2026.

[6] NIST National Cybersecurity Center of Excellence. "Summary of Comments: Software and AI Agent Identity and Authorization." NCCoE AI Identity Resource Hub, 2026.

[7] Rose, S., Borchert, O., Mitchell, S., and Connelly, S. "Zero Trust Architecture (SP 800-207)." NIST, August 2020.

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

[9] Fett, D., Campbell, B., Bradley, J., Lodderstedt, T., Jones, M., and Waite, D. "OAuth 2.0 Demonstrating Proof of Possession (DPoP) (RFC 9449)." IETF, September 2023.

[10] Campbell, B., Bradley, J., Sakimura, N., and Lodderstedt, T. "OAuth 2.0 Mutual-TLS Client Authentication and Certificate-Bound Access Tokens (RFC 8705)." IETF, February 2020.

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

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

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

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

[14] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." CSA, June 2026.