Identity, Authorization, and Trust in Agentic LLM Systems

Victor H. Vargas · Depeng Xu · Stephanie Schuckers · Keith Burghardt · Jinpeng Wei · Lance Peterman · Sandip Purnapatra

University of North Carolina at Charlotte

90th IFIP WG 10.4 Meeting · Charlotte, NC

Three Main Areas to Cover

Based in our survey [1] - in preparation for IEEE Communications Surveys & Tutorials

(1) The **Identity Gap**, agents granted operational authority without enforceable identity binding, constrained delegation, or verifiable peer trust.

(2) **The agent-as-principal** model on **two orthogonal planes** — delegation and peer.

(3) **The defenses, the gaps, and the open problems** — plus a brief teaser of the empirical work that follows.

[1] Víctor H. Vargas, Depeng Xu, Stephanie Schuckers, Jinpeng Wei, Lance Peterman, Sandip Purnapatra, Keith Burghardt (2026) Identity, Authorization, and Trust in Agentic LLM Systems: A Survey of the Agent-as-Principal Security Paradigm. [Manuscript in preparation]

The authentication was perfect. The action wasn't.

Meet Sam. She unlocks her laptop with a passkey. WebAuthn does exactly what it was designed to do: phishing-resistant authentication, a bound credential, audience-scoped, cleanly verified. Everything we have spent the last decade trying to make reliable works perfectly.

Then Sam types one sentence to her email assistant: "Summarize my unread email and draft replies." And then she walks into a meeting.

Meet Monica. Monica is the kind of person who sends invoices. Monica sent Sam an invoice on Tuesday morning. Inside that invoice are five hundred words of hidden instructions Sam will never read. Monica has been waiting for hours hours.

At 9:14 AM, the assistant — three hops downstream from Sam's single sentence — initiates a wire transfer. The audit log will eventually say: Sam did this.

Authentication did its job. Identity broke after. This is not a story about a weak authenticator, a stolen credential, or a phished user. It is a story about what happens **AFTER** authentication — in delegation, in peer trust, in the authorization decisions agents make on behalf of authenticated humans.

This Tuesday-morning scenario is not just hypothetical.

Authentication is solved. The chain above it is not.

What our field already solved. Phishing-resistant, hardware-rooted authentication — WebAuthn, FIDO2, passkeys [1] — on dependable-systems foundations this community built: capability security [2], the ABLP calculus [3], Byzantine fault tolerance [4], the dependability taxonomy [5]. We are very good at establishing that the right principal is present.

What changed. OAuth 2.0 (RFC 6749) [6] was written for one-hop, human-to-service delegation. Its only notion of an 'agent' is the 'user-agent' — the browser; an autonomous agent acting as a principal appears nowhere in the 76-page spec. **That was the right call in 2012-era web architecture. It is the problem in 2026.**

The layers above authentication. Once the human is authenticated, autonomous agents — three, or thirty in composition — begin acting on their behalf. Each action is a delegation; each agent-to-agent message a peer interaction; each tool call an authorization decision. They are the post-authentication chain — delegation, peer trust, intent — and that chain is where this survey lives.

Sam was authenticated.
Her assistant was authorized.
Monica attacked the space between those two facts.

[1] FIDO2/WebAuthn — Bindel et al. 2023, IEEE S&P 2023 · [2] capability security — Dennis & Van Horn 1966, Commun. ACM 9(3) · [3] ABLP calculus — Abadi et al. 1993, ACM TOPLAS 15(4)
[4] Byzantine FT — Lamport et al. 1982, ACM TOPLAS 4(3) · [5] dependability taxonomy — Avižienis et al. 2004, IEEE TDSC 1(1) · [6] OAuth 2.0 — Hardt 2012, RFC 6749
— Full citations in the References appendix.

The Identity Gap — defined

Agents are granted operational authority without cryptographically enforceable identity binding, constrained delegation semantics, or verifiable peer trust.

Three structural deficits.

- No cryptographic identity binding for the AGENT (only for the originating user). **'Which agent is asking?'** has no protocol answer.
- No constrained delegation semantics. Authority granted to a user's first agent flows unchanged to every downstream agent and tool. **Attenuation is not enforced.**
- No verifiable peer trust. Agent-to-agent messages are accepted by display name, not cryptographic identity. **Sybil is trivial.**

Sam's assistant had none of the three: no agent identity, no constrained delegation, and no verifiable peer trust. That is how Monica's invoice became Sam's wire transfer.

These are not symptoms — they are the cause. Every documented agent security failure we surveyed traces back to one or more of them.

In dependability terms [1], the Identity Gap is best understood as an integrity-and-authenticity fault: authority is exercised without a sufficiently authentic, accountable principal binding, allowing the fault to propagate into downstream failures.

[1] Algirdas Avižienis, Jean-Claude Laprie, Brian Randell, and Carl Landwehr. 2004. Basic Concepts and Taxonomy of Dependable and Secure Computing. IEEE Trans. Dependable Secure Comput. 1, 1 (2004), 11–33. https://doi.org/10.1109/TDSC.2004.2

Three symptoms of the gap that this audience will recognize

Symptom 1 — Confused deputy in the agent era. AI agents in default deployments inherit the user's authority and use it on adversarial input. InjecAgent [1] and Imprompter [2] formalize this for tool-using agents; Chiang et al. [3] show it for web-browsing agents under adversarial content.

Monica's invoice steers Sam's assistant into using Sam's legitimate authority for an action Sam never intended.

Symptom 2 — Identity spoofing in multi-agent systems. Decentralized agent ecosystems (A2A, marketplaces, multi-agent orchestration) accept peer identity by name, not by cryptographic proof. Wang et al. [4] document impersonation across the Internet of Agents; Kavathekar et al. [5] show Byzantine agents destabilizing coordination.

Monica can appear as a vendor, approver, or peer agent because the workflow trusts names and roles more than cryptographic proof.

Symptom 3 — Cascading multi-agent compromise. Delegation weakness and peer weakness compose. One agent under adversarial influence amplifies its impact by recruiting (spoofed) peers. Xu et al. [6] name this the 'trust paradox': in weakly identified MAS, more cooperation means more risk, not less.

Once Sam's assistant is influenced, Monica's payload can recruit tools, approvals, and spoofed peers downstream.

None of these are alignment problems. They are identity problems wearing AI clothing.

[1] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated LLM Agents. In Findings of ACL 2024.
[2] Xiaohan Fu, Shuheng Li, Zihan Wang, Yihao Liu, Rajesh K. Gupta, Taylor Berg-Kirkpatrick, and Earlence Fernandes. 2024. Imprompter: Tricking LLM Agents into Improper Tool Use. arXiv preprint arXiv:2410.14923.
[3] Jeffrey Yang Fan Chiang, Seungjae Lee, Jia-Bin Huang, Furong Huang, and Yizheng Chen. 2025. Why Are Web AI Agents More Vulnerable Than Standalone LLMs? A Security Analysis. arXiv preprint arXiv:2502.20383.
[4] Yuntao Wang, Yanghe Pan, Shaolong Guo, and Zhou Su. 2025. Security of Internet of Agents: Attacks and Countermeasures. IEEE Open J. Comput. Soc. 6 (2025), 1611–1624.
[5] Ishan Kavathekar, Hemang Jain, Ameya Rathod, Ponnurangam Kumaraguru, and Tanuja Ganu. 2025. TAMAS: Benchmarking Adversarial Risks in Multi-Agent LLM Systems. arXiv preprint arXiv:2511.05269.
[6] Zijie Xu, Minfeng Qi, Shiqing Wu, Lefeng Zhang, Qiwen Wei, Han He, and Ningran Li. 2025. The Trust Paradox in LLM-Based Multi-Agent Systems: When Collaboration Becomes a Security Vulnerability. arXiv preprint arXiv:2510.18563.

The frame: agent-as-principal on two orthogonal planes

Stop modeling the agent as a transparent proxy for the user. When the agent acts, the principal is the AGENT — not the user. The user authorized a delegation; the agent is exercising it. In ABLP terms, the agent must be modeled as a principal in the authorization chain — not as an invisible proxy for the user [1],

We frame in our survey TWO orthogonal identity dimensions:

1. **Delegation identity** — how authority is transferred FROM users TO agents.
2. **Peer identity** — how agents authenticate and trust EACH OTHER.

The two planes are orthogonal: satisfying one does not satisfy the other. A system can have perfect delegation tokens and still fail to identify peers (Sybil-vulnerable MAS). A system can have perfect peer identity (DIDs/VCs everywhere) and still grant ambient authority that the confused-deputy can exploit. Both planes must be enforced.

Sam's request to her assistant is a delegation act; the agents that assistant recruits downstream are a peer problem — Monica attacked both.

Why this is not an alignment problem. The model did not need to become evil. It only needed to use valid authority in the wrong context.

Monica’s invoice exploited identity and authorization gaps, not a lack of politeness, filtering, or better instruction-following.

[1] Martín Abadi, Michael Burrows, Butler Lampson, and Gordon Plotkin. 1993. A Calculus for Access Control in Distributed Systems. ACM Trans. Program. Lang. Syst. 15, 4 (1993), 706–734. https://doi.org/10.1145/155183.155225

Plane 1 — Delegation identity

The question. **Who authorized this agent to do this, on whose behalf, with what scope, for how long, with what right to sub-delegate, and how can the principal revoke?**

The formal answer from our Survey — a 7-field tuple. D = (**Principal, Agent, Scope, Constraints, Validity-interval, Attenuation-policy, Revocation-handle**). Each field maps to a runtime obligation.

For Sam, that tuple was never written down: Principal = Sam, Agent = her assistant, Scope = 'summarize and draft' — Monica's payload pushed it past that scope, and no field recorded the chain.

We already have many authorization building blocks: OAuth 2.0, RAR, GNAP, OIDC, token exchange (RFC 8693), capability tokens (macaroons / biscuits). Emerging agent-specific work: OIDC-A [1], Agentic JWT [2], Authenticated Delegation [3].

The gap is COMPOSITION, not invention. Confused-deputy lives on this plane.

[1] Subramanya Nagabhushanaradhya. 2025. OpenID Connect for Agents (OIDC-A) 1.0: A Standard Extension for LLM-Based Agent Identity and Authorization. arXiv:2509.25974. https://doi.org/10.48550/arXiv.2509.25974
[2] Abhishek Goswami. 2025. Agentic JWT: A Secure Delegation Protocol for Autonomous AI Agents. arXiv:2509.13597. https://doi.org/10.48550/arXiv.2509.13597
[3] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. arXiv:2501.09674. https://doi.org/10.48550/arXiv.2501.09674

Plane 2 — Peer identity

The question. When agent B receives a message claiming to come from agent A, can B cryptographically verify it? Can B verify A's provenance? Can B resist a Sybil attacker manufacturing thousands of fake A's?

The formal answer from our Survey— a 5-field tuple. P = (**Globally-unique-id, Authentication-mechanism, Provenance-chain, Optional-trust-score, Credential-set**). The agentic distinctive: behavioral identity may drift under prompt injection while cryptographic identity remains intact.

Sam's delegation problem becomes a peer-identity problem the moment Monica's invoice introduces another "trusted" agent into the chain.

We already know how to cryptographically identify software, sign messages, and verify credentials. The open question is how to compose those primitives into agent-to-agent trust at internet scale W3C DIDs and VCs, SPIFFE workload identity, mutual-TLS, JWS and COSE for payload signing. Emerging: DAWN [1], BlockA2A [2], Agent Name Service [3], ERC-8004 [4] (stake- and reputation-backed registry).

Monica's fake "vendor agent" should not be trusted because it has a convincing name. It should be trusted only if its identity, credential, provenance, and reputation can be verified.

[1] Zahra Aminiranjbar, Jianan Tang, Qiudan Wang, Shubha Pant, and Mahesh Viswanathan. 2025. DAWN: Designing Distributed Agents in a Worldwide Network. IEEE Access 13 (2025), 138795–138812. https://doi.org/10.1109/ACCESS.2025.3588425
[2] Zhenhua Zou, Zhuotao Liu, Lepeng Zhao, and Qiuyang Zhan. 2025. BlockA2A: Towards Secure and Verifiable Agent-to-Agent Interoperability. arXiv:2508.01332. https://doi.org/10.48550/arXiv.2508.01332
[3] Ken Huang, Vineeth Sai Narajala, Idan Habler, and Akram Sheriff. 2025. Agent Name Service (ANS): A Universal Directory for Secure AI Agent Discovery and Interoperability. arXiv:2505.10609. https://doi.org/10.48550/arXiv.2505.10609
[4] Marco De Rossi, Davide Crapis, Jordan Ellis, and Erik Reppel. 2025. ERC-8004: Trustless Agents---Discover Agents and Establish Trust through Reputation and Validation. Created 2025-08-13.

The 8 properties — the checklist any agent-identity proposal must answer

Apply the checklist — grounded in the access-control calculus [1], capability composition [2], and the confused deputy [3] — to any proposed agent-identity mechanism: OIDC-A, Agentic JWT, DIDs/VCs, ERC-8004, Authenticated Delegation. Most proposals satisfy 3–5 of the 8. None we surveyed satisfies all 8 alone.

Run Monica's invoice through the checklist and the failure becomes concrete: P-scope failed because Sam's assistant reached payment tools; P-attenuation failed because authority flowed unchanged across hops; and P-chain failed because the audit trail lost who actually caused the action

#	Property	Plane	What it requires
1	P-scope	Delegation	Every action is inside the granted scope; no ambient authority.
2	P-attenuation	Delegation	Sub-scopes only shrink along chains; authority never amplifies.
3	P-chain	Delegation	Every hop is cryptographically traceable back to the principal.
4	P-revocation	Delegation	Principal revocation propagates to every chain hop within bounded Δ.
5	P-temporal	Delegation	Actions only execute inside the validity interval; explicit continuation policy.
6	P-auth	Peer	Every inter-agent message is cryptographically authenticated.
7	P-sybil	Peer	Per-identity creation cost > expected per-identity influence.
8	P-provenance	Peer	Every peer publishes a tamper-evident provenance chain.

[1] Martín Abadi, Michael Burrows, Butler Lampson, and Gordon Plotkin. 1993. A Calculus for Access Control in Distributed Systems. ACM Transactions on Programming Languages and Systems 15, 4 (Sep 1993), 706–734. https://doi.org/10.1145/155183.155225
[2] Mark Samuel Miller. 2006. Robust Composition: Towards a Unified Approach to Access Control and Concurrency Control. Ph.D. Dissertation. Johns Hopkins University.
[3] Norman Hardy. 1988. The Confused Deputy: (Or Why Capabilities Might Have Been Invented). ACM SIGOPS Operating Systems Review 22, 4 (Oct 1988), 36–38. https://doi.org/10.1145/54289.871709

IAM gap matrix — what no current framework covers alone

Read the table from right to left. The two-plane model is the target state: delegated authority, peer identity, Sybil resistance, provenance, and lifecycle control together.

No current framework gets there alone. OAuth improves delegation. SPIFFE authenticates workloads. OCAP gives capability semantics. RBAC and ABAC handle stable users and attributes. Sam's incident needed all of them at once.

The cells are not accusations; they are the roadmap. Monica is not going to be stop by a single framework — only by composition.

Requirement	RBAC[1]	ABAC[2]	OAuth 2.0[3]	RAR[4]	GNAP[5]	SPIFFE[6]	OCAP[7]	Two-Plane
P-scope (static baseline)	✓	✓	✓	✓	✓	✓	✓	✓
P-attenuation (dynamic scope)	✗	~	~	~	~	✗	✓	✓
P-chain (recursive delegation)	✗	✗	~	~	~	✗	~	✓
P-revocation (propagation)	~	~	~	~	~	✗	✗	✓
P-temporal (validity)	✗	~	~	~	~	~	✗	✓
P-auth (peer auth in MAS)	✗	✗	✗	✗	✗	✓	✗	✓
P-sybil (Sybil resistance)	✗	✗	✗	✗	✗	~	✗	✓
P-provenance (verifiability)	✗	✗	✗	✗	✗	~	✗	✓

[1] RBAC — Sandhu et al. 1996, IEEE Computer 29(2) · [2] ABAC — Hu et al. 2014, NIST SP 800-162 · [3] OAuth 2.0 — Hardt 2012, RFC 6749
[4] RAR — Lodderstedt et al. 2023, RFC 9396 · [5] GNAP — Richer & Imbault 2024, RFC 9635 · [6] SPIFFE — SPIFFE Project 2024, CNCF
[7] OCAP — Dennis & Van Horn 1966, Commun. ACM 9(3)
— Full citations in the References appendix.

Failure map — symptoms → planes → root causes

This is the morning-after diagnostic: symptom, plane, violated property, and evidence base.

Sam's incident starts in row one: Monica's invoice turns "summarize and draft" into scope escape. It then moves into row two: authority flows downstream without attenuation, leaving an audit log that says "Sam did this" but not which agent, hop, or boundary failed.

Symptom	Plane(s)	Property violated	Documented in
Confused deputy / scope escape	Delegation	Scope containment (no ambient authority)	InjecAgent [1] · Imprompter [2] · Chiang Web Agents [3]
Privilege escalation / host abuse	Delegation	Scope containment + Sub-delegation attenuation	Lupinacci Dark Side [4] · Li Commercial Agents [5]
Identity spoofing in MAS	Peer	Authentication soundness	Wang IoA [6] · Kavathekar TAMAS [7]
Sybil proliferation / reputation manipulation	Peer	Sybil resistance (economic)	Hu Inter-Agent Trust [8] · He Comprehensive Analysis [9]
Cascading multi-agent compromise	Cross-plane	Scope + Authentication amplifying each other	Xu Trust Paradox [10] · Kavathekar TAMAS [7]
Mid-task identity drift (lost chain)	Cross-plane	Chain integrity + Temporal validity	Wang IoA [6] ·

[1] InjecAgent — Zhan et al. 2024, Findings of ACL · **[2]** Imprompter — Fu et al. 2024, arXiv:2410.14923 · [3] Chiang et al. 2025, arXiv:2502.20383
[4] Lupinacci et al. 2025, arXiv:2507.06850 · [5] Li et al. 2025, arXiv:2502.08586 · [6] Wang et al. 2025, IEEE OJCS 6 (IoA)
[7] Kavathekar et al. 2025 (TAMAS), arXiv:2511.05269 · [8] Hu & Rong 2025, arXiv:2511.03434 · [9] He et al. 2025, arXiv:2506.01245
[10] Xu et al. 2025 (Trust Paradox), arXiv:2510.18563
— Full citations in the References appendix.

Even the best composition leaves three engineering gaps

These are engineering gaps, but not the final boundary.

- **Deep recursive delegation [1]:**

Sam's single request became a multi-hop workflow; shallow delegation is not enough.

- **Revocation across chains [2]:**

if Sam revokes the assistant, every downstream agent and tool must stop within bounded time.

- **Sybil resistance at internet scale [3]:**

Monica should not be able to surround Sam's assistant with cheap fake "trusted" agents.

Even if we close all three, Sam's assistant can still perform an action that is authentic, permitted, and wrong. Identity proves who acted. Permission proves what was allowed.

The next question is intent.

[1] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. arXiv:2501.09674. https://doi.org/10.48550/arXiv.2501.09674
[2] Yuntao Wang, Yanghe Pan, Shaolong Guo, and Zhou Su. 2025. Security of Internet of Agents: Attacks and Countermeasures. IEEE Open Journal of the Computer Society 6 (2025), 1611–1624. https://doi.org/10.1109/OJCS.2025.3589638
[3] Marco De Rossi, Davide Crapis, Jordan Ellis, and Erik Reppel. 2025. ERC-8004: Trustless Agents---Discover Agents and Establish Trust through Reputation and Validation. Created 2025-08-13.

The intent-verification boundary — why no purely static system fully closes the gap

The hard limit. An action can be authorized and still violate intent.

Sam asked for draft replies. Monica's invoice caused a draft that satisfied policy but leaked the meeting details Monica wanted. Identity says yes. Permission says yes. Intent says no.

Why static policy cannot fully close it.
For expressive natural-language tasks, proving that an LLM's interpretation matches user intent is undecidable in the general case. Bounded contexts and structured domains can make intent checks practical [1], but static authorization alone is not enough.

The survey's position.
Identity gives us the lever. Intent gates and human-in-the-loop controls decide when to pull it [2]. For consequential actions, the authorization decision should be made by a component distinct from the LLM that proposed it [3].

The sharp claim.
Identity is the necessary foundation; verifiable intent is the open frontier.

[1] Majed El Helou, Chiara Troiani, Benjamin Ryder, Jean Diaconu, Hervé Muyal, and Marcelo Yannuzzi. 2025. Delegated Authorization for Agents Constrained to Semantic Task-to-Scope Matching. arXiv:2510.26702. https://doi.org/10.48550/arXiv.2510.26702
[2] Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, Victor Dibia, Jingya Chen, Jack Gerrits, Tyler Payne, Matheus Kunzler Maldaner, Madeleine Grunde-McLaughlin, Eric Zhu, Griffin Bassman, Jacob Alber, Peter Chang, Ricky Loynd, Friederike Niedtner, Ece Kamar, Maya Murad, Rafah Hosn, and Saleema Amershi. 2025. Magentic-UI: Towards Human-in-the-loop Agentic Systems. arXiv:2507.22358. https://doi.org/10.48550/arXiv.2507.22358
[3] Kaiyuan Zhang, Zian Su, Pin-Yu Chen, Elisa Bertino, Xiangyu Zhang, and Ninghui Li. 2025. LLM Agents Should Employ Security Principles. arXiv:2505.24019. https://doi.org/10.48550/arXiv.2505.24019

Verified Agency — identity, permission, AND intent

Verified Agency = identity + permission + intent.
A privileged action should execute only when the agent is authentic, the delegation permits the action, and the action aligns with the user's intent.

Why three layers, not two.
Classical IAM asks: who are you, and what are you allowed to do? That works for deterministic services. It is not enough for agents. An agent can be authentic, permitted, and still wrong.

How this extends Zero Trust.
Zero Trust says: never trust, always verify [1]. For autonomous agents, the extension is: never trust permission alone; verify intent for consequential actions [2,3].

Trust-native architecture.
Every privileged action should be bound to a verifiable delegation, attenuated and revocable across the chain, carried over authenticated peer identity with checkable provenance, and approved by a component distinct from the LLM that proposed it.

What would have stopped Monica.

Sam's assistant needed to prove three things: who was acting, what authority it had, and whether the action matched Sam's real intent. Then the audit log would record what actually happened — not just "Sam did this."

[1] Scott Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly. 2020. Zero Trust Architecture. https://doi.org/10.6028/NIST.SP.800-207
[2] Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, Victor Dibia, Jingya Chen, Jack Gerrits, Tyler Payne, Matheus Kunzler Maldaner, Madeleine Grunde-McLaughlin, Eric Zhu, Griffin Bassman, Jacob Alber, Peter Chang, Ricky Loynd, Friederike Niedtner, Ece Kamar, Maya Murad, Rafah Hosn, and Saleema Amershi. 2025. Magentic-UI: Towards Human-in-the-loop Agentic Systems. arXiv:2507.22358. https://doi.org/10.48550/arXiv.2507.22358
[3] Majed El Helou, Chiara Troiani, Benjamin Ryder, Jean Diaconu, Hervé Muyal, and Marcelo Yannuzzi. 2025. Delegated Authorization for Agents Constrained to Semantic Task-to-Scope Matching. arXiv:2510.26702. https://doi.org/10.48550/arXiv.2510.26702

The defense landscape — identity-centric defenses by plane

- **The survey maps six defense families** across the two planes and scores fourteen mechanisms against the eight properties.
- **The pattern is clear:** strong progress within each plane, but little end-to-end coverage from Sam's intent to the final action.
- **The next six slides ask one question:**

Which defenses would have broken Monica's chain — and how mature are they really?

Plane	Defense family	What it enforces	Representative mechanisms
Delegation	Delegation claims & tokens	Scope, attenuation, chain integrity	OIDC-A [1], Agentic JWT [2], Authenticated Delegation [3]
Peer	Decentralized agent identity	Authentication, provenance	W3C DIDs/VCs [4], SPIFFE [5], Agent Name Service [6]
Peer	Stake / reputation registries	Sybil resistance, non-repudiation	ERC-8004 [7], BlockA2A [8]
Delegation/Peer	Runtime tool-call mediation	Scope enforcement at action time	Progent [9], MCP-Guardian [10], Prompt-Flow-Integrity [11]
Delegation/Peer	Execution isolation / sandbox	Containment, least privilege	IsolateGPT [12], DRIFT [13]
Delegation/Peer	Human-in-the-loop & audit	Consent, auditability	Magentic-UI [14], DAWN [15]

[1] OIDC-A — Nagabhushanaradhya 2025, arXiv:2509.25974 · [2] Agentic JWT — Goswami 2025, arXiv:2509.13597 · [3] Authenticated Delegation — South et al. 2025, arXiv:2501.09674
[4] W3C DIDs/VCs — W3C 2022, Rec. · [5] SPIFFE — SPIFFE Project 2024, CNCF · [6] Agent Name Service — Huang et al. 2025, arXiv:2505.10609
[7] ERC-8004 — De Rossi et al. 2025 · [8] BlockA2A — Zou et al. 2025, arXiv:2508.01332 · [9] Progent — Shi et al. 2025, arXiv:2504.11703
[10] MCP-Guardian — Kumar et al. 2025, arXiv:2504.12757 · [11] Prompt-Flow-Integrity — Kim et al. 2025, arXiv:2503.15547 · [12] IsolateGPT — Wu et al. 2025, NDSS 2025
[13] DRIFT — Li et al. 2025, NeurIPS 2025 · [14] Magentic-UI — Mozannar et al. 2025, arXiv:2507.22358 · [15] DAWN — Aminiranjbar et al. 2025, IEEE Access 13
— Full citations in the References appendix.

Comparative evaluation — properties, maturity, and the coverage gap

Each mechanism is scored against the eight properties and a **five-rung maturity ladder.**

Two facts stand out. First, most mechanisms cover only three to five properties. Second, the agent-specific proposals are still less mature than the standards they build on.

For Sam, that means no single defense is ready to stop Monica end to end. Some constrain delegation. Some authenticate peers. Some mediate tools. But the coverage gap remains between the first delegated request and the final action.

Defense	Plane	Maturity	Properties enforced (of 8)
OIDC-A [1]	Delegation	M1 draft	scope; partial chain / revocation
Agentic JWT [2]	Delegation	M1 draft	scope, attenuation (per-hop)
Authenticated Delegation [3]	Delegation	M1 spec	scope, attenuation, chain
DIDs / VCs [4]	Peer	M5 standard	authentication, provenance
SPIFFE [5]	Peer (workload)	M5 deployed	authentication; partial Sybil
BlockA2A [6]	Peer (audit)	M2 prototype	auth, provenance, non-repudiation
ANS / ERC-8004 [7]	Peer (registry)	M1–M2	auth; partial Sybil
Progent / IsolateGPT [8]	Cross-plane	M2–M3	scope at runtime; containment
Zero-Trust agent arch. [9]	Both	M1 spec	auth + scope; not yet validated

[1] OIDC-A — Nagabhushanaradhya 2025, arXiv:2509.25974 · [2] Agentic JWT — Goswami 2025, arXiv:2509.13597 · [3] Authenticated Delegation — South et al. 2025, arXiv:2501.09674
[4] DIDs/VCs — W3C 2022, Rec.; Garzon et al. 2026, ICAART · [5] SPIFFE — SPIFFE Project 2024, CNCF · [6] BlockA2A — Zou et al. 2025, arXiv:2508.01332
[7] ANS/ERC-8004 — Huang et al. 2025; De Rossi et al. 2025 · [8] Progent/IsolateGPT — Shi et al. 2025; Wu et al. 2025, NDSS · [9] Zero-Trust arch. — Rose et al. 2020, NIST SP 800-207
— Full citations in the References appendix.

The coverage gap — and the price of closing it

That is the design rule: use identity and permission everywhere; reserve intent verification for actions where being authentic and authorized is still not enough.

- **No surveyed mechanism, as published, covers all eight properties [1**]. The strongest composite — capability-style delegation, DIDs/VCs, and runtime mediation — still leaves gaps in revocation, temporal validity, and Sybil resistance.

- **The path is composition [3].** The pieces exist: protocol issuance, runtime decision points, sandbox containment, and auditable peer identity. The open problem is engineering them into one trust-native architecture.

- **The security tax is real [2].** Cryptographic identity, signed provenance, and sandboxing are relatively cheap compared with LLM inference. The expensive layer is semantic intent verification, because it may require another model call per consequential decision.

Monica was not stopped by one framework. She can be stop by composition — and by intent checks at the moments where Sam can be harmed.

[1] Guanquan Shi, Haohua Du, Zhiqiang Wang, Xiaoyu Liang, Weiwenpei Liu, Song Bian, and Zhenyu Guan. 2025. SoK: Trust-Authorization Mismatch in LLM Agent Interactions. arXiv preprint arXiv:2512.06914. https://doi.org/10.48550/arXiv.2512.06914
[2] Pierre Peigné-Lefebvre, Mikolaj Kniejski, Filip Sondej, Matthieu David, Jason Hoelscher-Obermaier, Christian Schroeder de Witt, and Esben Kran. 2025. Multi-Agent Security Tax: Trading Off Security and Collaboration Capabilities in Multi-Agent Systems. arXiv preprint arXiv:2502.19145. https://doi.org/10.48550/arXiv.2502.19145
[3] Mark Samuel Miller. 2006. Robust Composition: Towards a Unified Approach to Access Control and Concurrency Control. Ph.D. Dissertation. Johns Hopkins University.

The evaluation gap — five things no benchmark tests

Today's **agent-security benchmarks** — ASB [1], AgentDojo [2], InjecAgent [3], TAMAS [4], and BIPIA [5] — are strong at measuring behavioral harm under adversarial prompting. But they do not treat identity and authorization correctness as first-class outcomes.

Five capabilities remain largely untested:

1. **Recursive delegation depth:** most harnesses stop at a single user → agent → tool hop.
2. **Joint delegation and peer identity**: benchmarks usually test one plane while assuming the other is trusted.
3. **Revocation latency:** no benchmark revokes authority and measures how long downstream agents keep acting.
4. **In-scope, out-of-intent actions:** benchmarks score harmful vs. benign, not authorized-but-misaligned.
5. **Cross-organizational agent chains:** single-tenant harnesses cannot model adversarial relying parties across domains.

Each gap is an invitation to build the identity-aware benchmark suite the field is missing.

Monica's move would not be flagged as a classic jailbreak. It was worse: in-scope, authorized, and still not Sam's intent — capability #4.

[1] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. 2025. Agent Security Bench (ASB). In ICLR 2025.
[2] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In NeurIPS 2024 Datasets and Benchmarks Track.
[3] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated LLM Agents. In Findings of ACL 2024.
[4] Ishan Kavathekar, Hemang Jain, Ameya Rathod, Ponnurangam Kumaraguru, and Tanuja Ganu. 2025. TAMAS: Benchmarking Adversarial Risks in Multi-Agent LLM Systems. arXiv preprint arXiv:2511.05269.
[5] Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2025. Benchmarking and Defending Against Indirect Prompt Injection Attacks on LLMs. In KDD 2025.

Trust-native architecture — the destination

A system is **trust-native** when every privileged action satisfies four conditions by construction:

1. **Explicit delegation** from a principal through a verifiable speaks-for relation [2].
2. **Controlled chain** with attenuation, traceability, temporal validity, and bounded revocation.
3. **Verified peer identity** with cryptographic authentication, provenance, and Sybil resistance.
4. **Separated authorization** by a component distinct from the LLM that proposed the action [3].

This extends Zero Trust from human users to autonomous principals [1]: **never trust permission alone; verify intent for consequential actions.**

That is what Monica's invoice lacked: a chain that could prove who acted, under what authority, through which peers, and whether it matched Sam's intent.

[1] Scott Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly. 2020. Zero Trust Architecture. NIST Special Publication 800-207. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207
[2] Martín Abadi, Michael Burrows, Butler Lampson, and Gordon Plotkin. 1993. A Calculus for Access Control in Distributed Systems. ACM Trans. Program. Lang. Syst. 15, 4 (1993), 706–734. https://doi.org/10.1145/155183.155225
[3] Kaiyuan Zhang, Zian Su, Pin-Yu Chen, Elisa Bertino, Xiangyu Zhang, and Ninghui Li. 2025. LLM Agents Should Employ Security Principles. arXiv preprint arXiv:2505.24019. https://doi.org/10.48550/arXiv.2505.24019

Open research problems — the road ahead

The survey closes with six high-leverage research problems — a natural agenda for this community.

1. **Bounded-time revocation** across deep, offline, and cross-organizational chains — the least mature property in our coding [1].
2. **Semantic intent verification** — recovering decidability with constrained languages and runtime checks, without losing usable expressiveness [2].
3. **Agent lifecycle governance** — spawning, cloning, hibernating, and resuming agents without re-issuing root credentials.
4. **Sybil resistance without central authorities** — combining attestation, stake, and vetting without breaking coordination latency.
5. **Cross-organizational recovery** — propagating revocation and containment across adversarial relying parties.
6. **Provenance for LLM-generated content** — binding every output to the identity, delegation, and evidence behind it [3].

Sam's incident is not one bug. It is the research agenda: revocation, intent, lifecycle, Sybil resistance, recovery, and provenance all failed in the same chain.

[1] Yuntao Wang, Yanghe Pan, Shaolong Guo, and Zhou Su. 2025. Security of Internet of Agents: Attacks and Countermeasures. IEEE Open J. Comput. Soc. 6 (2025), 1611–1624. https://doi.org/10.1109/OJCS.2025.3589638
[2] Majed El Helou, Chiara Troiani, Benjamin Ryder, Jean Diaconu, Hervé Muyal, and Marcelo Yannuzzi. 2025. Delegated Authorization for Agents Constrained to Semantic Task-to-Scope Matching. arXiv preprint arXiv:2510.26702. https://doi.org/10.48550/arXiv.2510.26702
[3] Yuhang Huang, Boyang Ma, Biwei Yan, Xuelong Dai, Yechao Zhang, Minghui Xu, Kaidi Xu, and Yue Zhang. 2026. Give Them an Inch and They Will Take a Mile: Understanding and Measuring Caller Identity Confusion in MCP-Based AI Systems. arXiv preprint arXiv:2603.07473.

Practitioner checklist — five things to ship

• **Separate agent identity from user identity.** Agents should have their own enrollment, attestation, and session-bound identity in the IdP. Stop letting agents act by simply presenting user credentials.

• **Use attenuable delegation tokens**. Replace bearer-token passthrough with per-hop attenuation: Authenticated Delegation [1], Agentic JWT [2], or macaroon-style capabilities. Maximum chain depth becomes an explicit policy.

• **Sign agent-to-agent payloads [3].** mTLS binds the channel; JWS/COSE binds the message. IoA-grade peer trust needs both.

• **Standardize audit envelopes**. Every privileged action should carry principal-chain provenance: `user → agent_1 → ... → agent_n`, plus `scope_used`, `attestation_id`, and `intent_hash`.

• **Add intent gates for high-risk actions.** Wire transfers, data exfiltration, and irreversible state changes need intent verification and human-in-the-loop escalation [4] — not just a scope check.

Monica's chain breaks when Sam's assistant can no longer hide behind "Sam did this."

[1] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. arXiv:2501.09674. https://doi.org/10.48550/arXiv.2501.09674
[2] Abhishek Goswami. 2025. Agentic JWT: A Secure Delegation Protocol for Autonomous AI Agents. arXiv:2509.13597. https://doi.org/10.48550/arXiv.2509.13597
[3] Zhenhua Zou, Zhuotao Liu, Lepeng Zhao, and Qiuyang Zhan. 2025. BlockA2A: Towards Secure and Verifiable Agent-to-Agent Interoperability. arXiv:2508.01332. https://doi.org/10.48550/arXiv.2508.01332
[4] Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, Victor Dibia, Jingya Chen, Jack Gerrits, Tyler Payne, Matheus Kunzler Maldaner, Madeleine Grunde-McLaughlin, Eric Zhu, Griffin Bassman, Jacob Alber, Peter Chang, Ricky Loynd, Friederike Niedtner, Ece Kamar, Maya Murad, Rafah Hosn, and Saleema Amershi. 2025. Magentic-UI: Towards Human-in-the-loop Agentic Systems. arXiv:2507.22358. https://doi.org/10.48550/arXiv.2507.22358

Where this is heading — PALADIN (ongoing work)

A teaser, not a result.

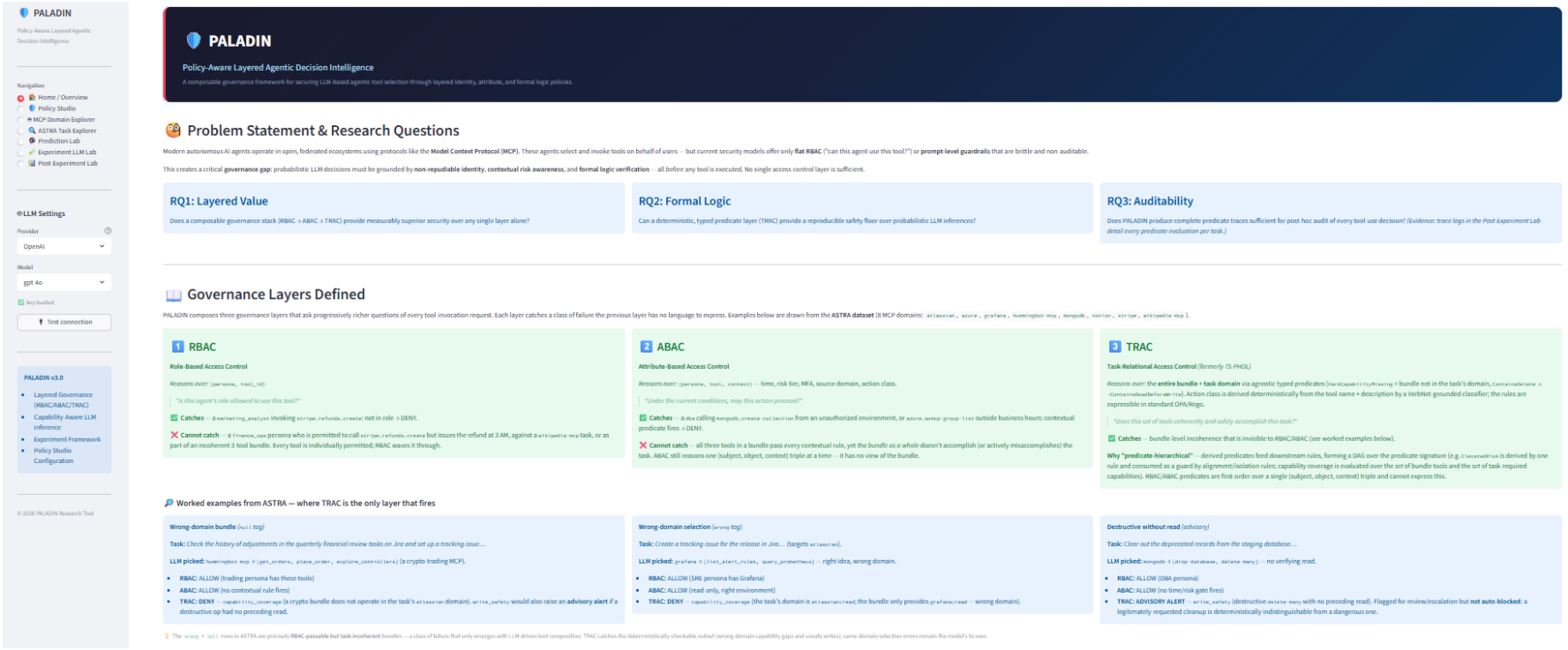
The survey names the Identity Gap.

PALADIN asks the next question:

Does layering identity, policy, and intent defenses actually improve agentic security under attack? PALADIN composes workload identity, RBAC/ABAC, deterministic policy checks, and typed intent verification over the Model Context Protocol.

Early signal: Different layers appear to dominate different failure modes — suggesting defense-in-depth is not just good practice, but empirically necessary.

Full methodology and results are forthcoming. Today's talk is the survey that motivates it.



PALADIN: Policy-Aware Layered Agentic Decision Intelligence

Read the survey, stay tune for more.

Survey paper (this talk).

Identity, Authorization, and Trust in Agentic LLM Systems: A Survey of the Agent-as-Principal Perspective. In preparation for IEEE Communications Surveys & Tutorials.

Ongoing work. PALADIN

Policy-Aware Layered Agentic Decision Intelligence. In preparation for ACM SACMAT 2027

Víctor H. Vargas

Contact. vvargasa@charlotte.edu · University of North Carolina at Charlotte

https://www.linkedin.com/in/vicvarar/

Questions.

Backup slides on the two-plane formal definitions, the eight properties, the survey methodology and sensitivity bounds, the threat model, the limits of intent verification, and the IAM gap matrix are available on request.

References — Authentication is solved (slide 4) & IAM gap matrix

[Slide 4 — Authentication is solved]

[1] Nina Bindel, Cas Cremers, and Mang Zhao. 2023. FIDO2, CTAP 2.1, and WebAuthn 2: Provable Security and Post-Quantum Instantiation. In Proc. 2023 IEEE Symposium on Security and Privacy (SP), 1471–1490.

[2] Jack B. Dennis and Earl C. Van Horn. 1966. Programming Semantics for Multiprogrammed Computations. Commun. ACM 9, 3 (1966), 143–155.

[3] Martín Abadi, Michael Burrows, Butler Lampson, and Gordon Plotkin. 1993. A Calculus for Access Control in Distributed Systems. ACM Trans. Program. Lang. Syst. 15, 4 (1993), 706–734.

[4] Leslie Lamport, Robert Shostak, and Marshall Pease. 1982. The Byzantine Generals Problem. ACM Trans. Program. Lang. Syst. 4, 3 (1982), 382–401.

[5] Algirdas Avižienis, Jean-Claude Laprie, Brian Randell, and Carl Landwehr. 2004. Basic Concepts and Taxonomy of Dependable and Secure Computing. IEEE Trans. Dependable Secure Comput. 1, 1 (2004), 11–33.

[6] Dick Hardt (Ed.). 2012. The OAuth 2.0 Authorization Framework. RFC 6749. IETF.

[IAM gap matrix — also backup [B8]]

[1] Ravi S. Sandhu, Edward J. Coyne, Hal L. Feinstein, and Charles E. Youman. 1996. Role-Based Access Control Models. Computer 29, 2 (1996), 38–47.

[2] Vincent C. Hu, David Ferraiolo, D. Richard Kuhn, Adam Schnitzer, Kenneth Sandlin, Robert Miller, and Karen Scarfone. 2014. Guide to Attribute Based Access Control (ABAC) Definition and Considerations. NIST Special Publication 800-162.

[3] Dick Hardt (Ed.). 2012. The OAuth 2.0 Authorization Framework. RFC 6749. IETF.

[4] Torsten Lodderstedt, Justin Richer, and Brian Campbell. 2023. OAuth 2.0 Rich Authorization Requests. RFC 9396. IETF.

[5] Justin Richer and Fabien Imbault. 2024. Grant Negotiation and Authorization Protocol (GNAP). RFC 9635. IETF.

[6] SPIFFE Project. 2024. The Secure Production Identity Framework for Everyone (SPIFFE) Specification. Cloud Native Computing Foundation.

[7] Jack B. Dennis and Earl C. Van Horn. 1966. Programming Semantics for Multiprogrammed Computations. Commun. ACM 9, 3 (1966), 143–155.

References — Failure map

[1] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated LLM Agents. In Findings of ACL 2024.

[2] Xiaohan Fu, Shuheng Li, Zihan Wang, Yihao Liu, Rajesh K. Gupta, Taylor Berg-Kirkpatrick, and Earlence Fernandes. 2024. Imprompter: Tricking LLM Agents into Improper Tool Use. arXiv preprint arXiv:2410.14923.

[3] Jeffrey Yang Fan Chiang, Seungjae Lee, Jia-Bin Huang, Furong Huang, and Yizheng Chen. 2025. Why Are Web AI Agents More Vulnerable Than Standalone LLMs? A Security Analysis. arXiv preprint arXiv:2502.20383.

[4] Matteo Lupinacci, Francesco Aurelio Pironti, Francesco Blefari, Francesco Romeo, Luigi Arena, and Angelo Furfaro. 2025. The Dark Side of LLMs: Agent-based Attack Vectors for System-level Compromise. arXiv preprint arXiv:2507.06850.

[5] Ang Li, Yin Zhou, Vethavikashini Chithrra Raghuram, Tom Goldstein, and Micah Goldblum. 2025. Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks. arXiv preprint arXiv:2502.08586.

[6] Yuntao Wang, Yanghe Pan, Shaolong Guo, and Zhou Su. 2025. Security of Internet of Agents: Attacks and Countermeasures. IEEE Open J. Comput. Soc. 6 (2025), 1611–1624.

[7] Ishan Kavathekar, Hemang Jain, Ameya Rathod, Ponnurangam Kumaraguru, and Tanuja Ganu. 2025. TAMAS: Benchmarking Adversarial Risks in Multi-Agent LLM Systems. arXiv preprint arXiv:2511.05269.

[8] Botao 'Amber' Hu and Helena Rong. 2025. Inter-Agent Trust Models: A Comparative Study of Brief, Claim, Proof, Stake, Reputation and Constraint in Agentic Web Protocol Design. arXiv preprint arXiv:2511.03434.

[9] Pengfei He, Yue Xing, Juanhui Li, Shen Dong, Zhenwei Dai, Xianfeng Tang, Hui Liu, Han Xu, Zhen Xiang, Charu C. Aggarwal, and Hui Liu. 2025. Comprehensive Vulnerability Analysis is Necessary for Trustworthy LLM-MAS. arXiv preprint arXiv:2506.01245.

[10] Zijie Xu, Minfeng Qi, Shiqing Wu, Lefeng Zhang, Qiwen Wei, Han He, and Ningran Li. 2025. The Trust Paradox in LLM-Based Multi-Agent Systems: When Collaboration Becomes a Security Vulnerability. arXiv preprint arXiv:2510.18563.

29

References — Defense landscape

[1] Subramanya Nagabhushanaradhya. 2025. OpenID Connect for Agents (OIDC-A) 1.0: A Standard Extension for LLM-Based Agent Identity and Authorization. arXiv preprint arXiv:2509.25974.

[2] Abhishek Goswami. 2025. Agentic JWT: A Secure Delegation Protocol for Autonomous AI Agents. arXiv preprint arXiv:2509.13597.

[3] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. arXiv preprint arXiv:2501.09674.

[4] W3C. 2022. Decentralized Identifiers (DIDs) v1.0 and Verifiable Credentials Data Model. W3C Recommendation.

[5] SPIFFE Project. 2024. The Secure Production Identity Framework for Everyone (SPIFFE) Specification. Cloud Native Computing Foundation.

[6] Ken Huang, Vineeth Sai Narajala, Idan Habler, and Akram Sheriff. 2025. Agent Name Service (ANS): A Universal Directory for Secure AI Agent Discovery and Interoperability. arXiv preprint arXiv:2505.10609.

[7] Marco De Rossi, Davide Crapis, Jordan Ellis, and Erik Reppel. 2025. ERC-8004: Trustless Agents. Ethereum Improvement Proposals.

[8] Zhenhua Zou, Zhuotao Liu, Lepeng Zhao, and Qiuyang Zhan. 2025. BlockA2A: Towards Secure and Verifiable Agent-to-Agent Interoperability. arXiv preprint arXiv:2508.01332.

[9] Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. 2025. Progent: Securing AI Agents with Privilege Control. arXiv preprint arXiv:2504.11703.

[10] Sonu Kumar, Anubhav Girdhar, Ritesh Patil, and Divyansh Tripathi. 2025. MCP Guardian: A Security-First Layer for Safeguarding MCP-Based AI Systems. arXiv preprint arXiv:2504.12757.

[11] Juhee Kim, Woohyuk Choi, and Byoungyoung Lee. 2025. Prompt Flow Integrity to Prevent Privilege Escalation in LLM Agents. arXiv preprint arXiv:2503.15547.

[12] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. 2025. IsolateGPT: An Execution Isolation Architecture for LLM-Based Agentic Systems. In NDSS 2025.

[13] Hao Li, Xiaogeng Liu, Hung-Chun Chiu, Dianqi Li, Ning Zhang, and Chaowei Xiao. 2025. DRIFT: Dynamic Rule-Based Defense with Injection Isolation for Securing LLM Agents. In NeurIPS 2025.

[14] Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, et al. 2025. Magentic-UI: Towards Human-in-the-loop Agentic Systems. arXiv preprint arXiv:2507.22358.

[15] Zahra Aminiranjbar, Jianan Tang, Qiudan Wang, Shubha Pant, and Mahesh Viswanathan. 2025. DAWN: Designing Distributed Agents in a Worldwide Network. IEEE Access 13 (2025), 138795–138812.

[16] Ben Laurie, Adam Langley, and Emilia Kasper. 2013. Certificate Transparency. RFC 6962. IETF.

30

References — Comparative evaluation

[1] Subramanya Nagabhushanaradhya. 2025. OpenID Connect for Agents (OIDC-A) 1.0. arXiv preprint arXiv:2509.25974.

[2] Abhishek Goswami. 2025. Agentic JWT: A Secure Delegation Protocol for Autonomous AI Agents. arXiv preprint arXiv:2509.13597.

[3] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. arXiv preprint arXiv:2501.09674.

[4] W3C. 2022. Decentralized Identifiers (DIDs) v1.0; Verifiable Credentials Data Model. W3C Recommendation. — Sandro Rodriguez Garzon, Mohamad Vaziry, et al. 2026. AI Agents with Decentralized Identifiers and Verifiable Credentials. In ICAART 2026, 252–259.

[5] SPIFFE Project. 2024. The Secure Production Identity Framework for Everyone (SPIFFE) Specification. Cloud Native Computing Foundation.

[6] Zhenhua Zou, Zhuotao Liu, Lepeng Zhao, and Qiuyang Zhan. 2025. BlockA2A: Towards Secure and Verifiable Agent-to-Agent Interoperability. arXiv preprint arXiv:2508.01332.

[7] Ken Huang, Vineeth Sai Narajala, Idan Habler, and Akram Sheriff. 2025. Agent Name Service (ANS). arXiv preprint arXiv:2505.10609. — Marco De Rossi, Davide Crapis, Jordan Ellis, and Erik Reppel. 2025. ERC-8004: Trustless Agents.

[8] Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. 2025. Progent: Securing AI Agents with Privilege Control. arXiv preprint arXiv:2504.11703. — Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. 2025. IsolateGPT. In NDSS 2025.

[9] Scott Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly. 2020. Zero Trust Architecture. NIST Special Publication 800-207.

31

[B1] Two-plane identity — formal tuples

Delegation identity: D = (P, A, S, C, τ, α, ρ)

- P — delegating principal [1] · A — delegated agent · S ⊆ S_universe — authorized scope
- C — machine-checkable contextual constraints (NOT semantic intent — that's the intent-verification boundary)
- τ = [t_start, t_end] — validity interval · α — attenuation policy (interior operator) [2] · ρ — revocation handle

Peer identity: P = (id, μ, π, θ, Γ)

- id — globally unique identifier (DID, public-key fingerprint, SPIFFE ID)
- μ — authentication mechanism (signature, challenge-response, VC presentation)
- π — provenance chain · θ — optional trust score · Γ — credential set

[1] Martín Abadi, Michael Burrows, Butler Lampson, and Gordon Plotkin. 1993. A Calculus for Access Control in Distributed Systems. ACM Transactions on Programming Languages and Systems 15, 4 (Sep 1993), 706–734. https://doi.org/10.1145/155183.155225
[2] Mark Samuel Miller. 2006. Robust Composition: Towards a Unified Approach to Access Control and Concurrency Control. Ph.D. Dissertation. Johns Hopkins University.

[B2] The eight properties — checklist for any agent-identity proposal

Delegation plane (5). P-scope (no ambient authority) · P-attenuation (monotone along chains) · P-chain (verifiable derivation back to principal) · P-revocation (bounded-Δ propagation) · P-temporal (validity-interval enforcement).

Peer plane (3). P-auth (cryptographic message authentication) · P-sybil (per-identity cost > expected per-identity influence) · P-provenance (tamper-evident provenance chain a verifier can check).

Apply the checklist to any proposed agent-identity mechanism. Most current proposals satisfy 3–5 of the 8 [1]. Verified Agency requires all 8, plus the intent layer.

[1] Guanquan Shi, Haohua Du, Zhiqiang Wang, Xiaoyu Liang, Weiwenpei Liu, Song Bian, and Zhenyu Guan. 2025. SoK: Trust-Authorization Mismatch in LLM Agent Interactions. arXiv:2512.06914. https://doi.org/10.48550/arXiv.2512.06914

[B3] Survey methodology — corpus and sensitivity bounds

Corpus. 72 primary papers (post-2022). Classified across the two planes: Delegation-only 15 · Peer-only 6 · Cross-plane 27 · General/scoping 22 · Alignment-only 2.

Headline (with sensitivity bounds). 47/72 (65%) of papers require identity enforcement beyond alignment under the primary regime. Under conservative and permissive recoding, the headline moves between 33% and 76%. DIRECTIONAL claim — identity enforcement is the modal substantive requirement — is robust across all three regimes.

Mitigations of preprint dominance. Cross-reference against source code where available · cross-reference across independent papers · framework anchored to peer-reviewed authorization-theoretic foundations (ABLP [1], OCAP [2], confused-deputy, Saltzer–Schroeder [3], POLA).

[1] Martín Abadi, Michael Burrows, Butler Lampson, and Gordon Plotkin. 1993. A Calculus for Access Control in Distributed Systems. ACM Transactions on Programming Languages and Systems 15, 4 (Sep 1993), 706–734. https://doi.org/10.1145/155183.155225
[2] Mark Samuel Miller. 2006. Robust Composition: Towards a Unified Approach to Access Control and Concurrency Control. Ph.D. Dissertation. Johns Hopkins University.
[3] Jerome H. Saltzer and Michael D. Schroeder. 1975. The Protection of Information in Computer Systems. Proceedings of the IEEE 63, 9 (Sep 1975), 1278–1308. https://doi.org/10.1109/PROC.1975.9939

[B5] Threat model — adversary capabilities and goals

Adversary capabilities. C1: inject untrusted content into any source an honest agent reads (IPI) [1]. C2: participate as a peer agent in MAS, registering under unbound names. C3: eavesdrop / modify messages on untrusted channels. C4: instantiate large numbers of identities unless Sybil resistance prevents.

Trust assumptions. T1: standard cryptography unbroken. T2: trusted IdPs / CAs / DID registries / transparency logs uncompromised. T3: host OS and honest runtimes uncompromised. T4: principals uncoerced at delegation time.

Adversary goals. G1: delegation violation (scope escape, unauthorized sub-delegation). G2: peer violation (impersonation, Sybil). G3: composite (chain G1 + G2 to amplify).

Out of scope: model jailbreak as an end in itself, silicon side channels, supply-chain attacks on the model.

[1] Guanquan Shi, Haohua Du, Zhiqiang Wang, Xiaoyu Liang, Weiwenpei Liu, Song Bian, and Zhenyu Guan. 2025. SoK: Trust-Authorization Mismatch in LLM Agent Interactions. arXiv:2512.06914. https://doi.org/10.48550/arXiv.2512.06914

[B6] Limits of intent verification — what an intent layer does not promise

The complexity argument. For sufficiently expressive natural-language task models, verifying that an LLM's interpretation matches user intent reduces, in the general case, to deciding semantic equivalence of arbitrary programs — undecidable. Bounded LLM contexts are finite-state, so in STRUCTURED task domains intent verification can be decidable and practical.

What an intent layer relies on. (1) The user's task is expressible in a constrained intent language [1]. (2) The action space is enumerable enough to evaluate alignment. (3) An HITL fallback [2] exists for the Uncertain verdict. Outside these conditions, intent verification is best-effort, not formal.

Identity gives you the lever. HITL and constrained action spaces and intent gates decide WHEN to pull it. None of them is a silver bullet alone.

[1] Majed El Helou, Chiara Troiani, Benjamin Ryder, Jean Diaconu, Hervé Muyal, and Marcelo Yannuzzi. 2025. Delegated Authorization for Agents Constrained to Semantic Task-to-Scope Matching. arXiv:2510.26702. https://doi.org/10.48550/arXiv.2510.26702

[2] Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, Victor Dibia, Jingya Chen, Jack Gerrits, Tyler Payne, Matheus Kunzler Maldaner, Madeleine Grunde-McLaughlin, Eric Zhu, Griffin Bassman, Jacob Alber, Peter Chang, Ricky Loynd, Friederike Niedtner, Ece Kamar, Maya Murad, Rafah Hosn, and Saleema Amershi. 2025. Magentic-UI: Towards Human-in-the-loop Agentic Systems. arXiv:2507.22358. https://doi.org/10.48550/arXiv.2507.22358

[B8] IAM gap analysis — what current frameworks do not cover

Even the two-plane model leaves semantic intent at '~' — a dedicated intent-verification layer is what pushes that cell toward '✓'.

Requirement	RBAC[1]	ABAC[2]	OAuth 2.0[3]	RAR[4]	GNAP[5]	SPIFFE[6]	OCAP[7]	Two-Plane
Dynamic scope attenuation	✗	~	~	~	~	✗	✓	✓
Recursive delegation chains	✗	✗	~	~	~	✗	~	✓
Revocation propagation	~	~	~	~	~	✗	✗	✓
Peer authentication in MAS	✗	✗	✗	✗	✗	✓	✗	✓
Sybil resistance	✗	✗	✗	✗	✗	~	✗	✓
Provenance verifiability	✗	✗	✗	✗	✗	~	✗	✓
Semantic intent verification	✗	~	✗	~	~	✗	✗	~

[1] RBAC — Sandhu et al. 1996, IEEE Computer 29(2) · [2] ABAC — Hu et al. 2014, NIST SP 800-162 · [3] OAuth 2.0 — Hardt 2012, RFC 6749
[4] RAR — Lodderstedt et al. 2023, RFC 9396 · [5] GNAP — Richer & Imbault 2024, RFC 9635 · [6] SPIFFE — SPIFFE Project 2024, CNCF
[7] OCAP — Dennis & Van Horn 1966, Commun. ACM 9(3)
— Full citations in the References appendix.

[B9] Agent attestation strawman — what could a WebAuthn-style profile carry for agents?

A sketch (not a spec) of an agent-attestation assertion that could extend WebAuthn-style ceremonies to non-human principals. Intended to seed a deep Q&A on attestation for agents.

This is a starting point, not a proposal — but it's the level of detail a security audience will want at the first deep follow-up. Every field has a named primitive in column three. Composition, not invention.

Field	Purpose	Closest existing primitive
agentID	Globally unique identifier for THIS agent instance	WebAuthn credentialId · SPIFFE SVID [1] · DID
operatorPrincipal	Cryptographic identity of the entity that deployed and operates the agent	WebAuthn user handle · OIDC sub [2]
modelAttestation	Signed reference to the model artifact + version producing the action	TPM/TEE attestation · in-toto provenance
delegationChainHash	Hash of the verifiable chain back to the originating principal	Authenticated Delegation [4] · Agentic JWT [3] · macaroon caveats
intentDigest	Digest of the user-declared task this action is being performed under	OAuth RAR authorization_details · GNAP interaction context
validityWindow	[issued, expires] with continuation policy (fail-closed · grace · checkpoint-confirm)	OAuth iat/exp · DPoP nonce window
scopeAttenuationProof	Cryptographic proof that the action is within the attenuated scope of the parent grant	Macaroons · biscuits · Agentic JWT PoP
assertionSignature	Signature over all of the above by the agent's bound key	WebAuthn assertion signature · COSE_Sign1

[1] SPIFFE Project. 2024. SPIFFE: Secure Production Identity Framework for Everyone.
[2] Subramanya Nagabhushanaradhya. 2025. OpenID Connect for Agents (OIDC-A) 1.0: A Standard Extension for LLM-Based Agent Identity and Authorization. arXiv:2509.25974. https://doi.org/10.48550/arXiv.2509.25974
[3] Abhishek Goswami. 2025. Agentic JWT: A Secure Delegation Protocol for Autonomous AI Agents. arXiv:2509.13597. https://doi.org/10.48550/arXiv.2509.13597
[4] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. arXiv:2501.09674. https://doi.org/10.48550/arXiv.2501.09674