From Implicit Trust to Continuous Verification: A Position on the Authorization of AI Agents in Production Services

Diego S. Araujo[1], **Silvio E. Quincozes**[2], **Rodrigo Brandão Mansilha**[3]

[1]AI Horizon Labs, PPGES, Universidade Federal do Pampa (UNIPAMPA)

{diegoaraujo.aluno, silvioquincozes, mansilha}@unipampa.edu.br

Abstract. *Production services increasingly receive requests from AI agents that act under delegated authority, live for minutes, and cross trust boundaries that were never mapped. We argue that admitting such requests safely requires abandoning implicit trust, so that every request is verified, every time. We describe an architecture of three conceptual layers that follow the life cycle of an agent request: cryptographic pre-authorization at the edge, workload identity attested by the execution platform, and contextual decisions externalized to a decoupled policy engine. We conclude with the field's main open problem: reconciling the delegation asserted by the organizational plane with the identity attested by the infrastructure.*

1. Introduction

Production services were designed for predictable human users or static software clients. AI agents break that assumption, since they act under delegated authority, live for only a few minutes, chain delegations, and cross heterogeneous trust boundaries [South et al. 2025a]. For this new consumer profile, classic IAM falls short in four respects. Static secrets invite leakage and replay when consumers are ephemeral; bearer credentials hide who delegated the access and with what scope [Tallam 2026]; coarse-grained scopes do not represent the intent of a concrete operation; and stable roles do not accommodate decisions that depend on attributes and context at every call. Our thesis is that **admitting AI agent requests in a production service requires replacing implicit trust with continuous verification, request by request**. In practice, the service then acts as a robust Policy Enforcement Point (PEP), which performs the cryptographic checks at the edge and delegates the semantic decision to a decoupled Policy Decision Point (PDP). The central gap, however, is not technological, since the cryptographic building blocks that are needed already exist and are standardized. It is conceptual: what is missing is an arrangement that composes those blocks at the point where the request is received.

2. An Architecture of Three Conceptual Layers

The architecture, summarized in Figure 1, organizes the five concepts into three layers: pre-authorization with proof of possession; attested identity and its binding; intent, evaluation of delegation, and externalized decision. The first layer establishes the cryptographic preconditions on the mandatory path of the service. The presented credential must be untampered, addressed to this service, and within its validity period, and the requester must prove possession of the key to which it is bound; bearer credentials are therefore inadmissible for agentic consumers. Requests that fail these checks are rejected at the edge, before any policy decision.

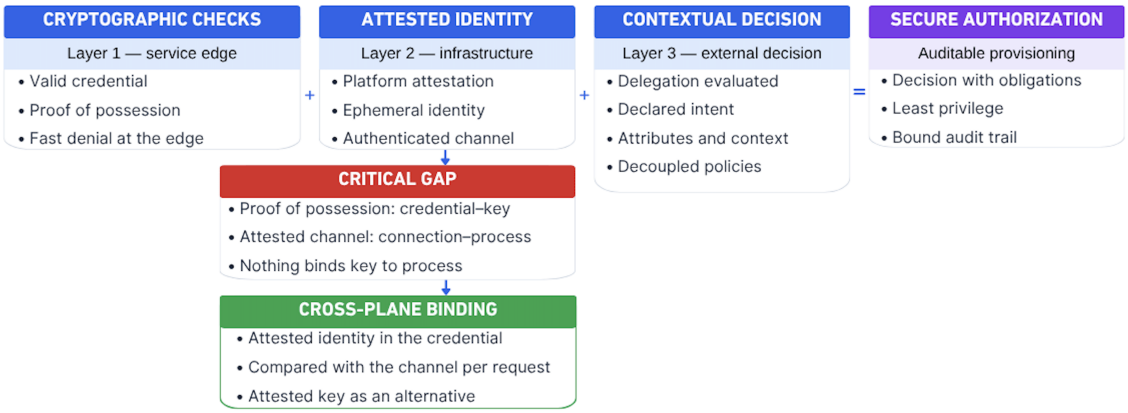

Figura 1. The three conceptual layers and the outcome of authorization. The critical gap and the binding that resolves it lie in Layer 2.

The second layer establishes who executes. The agent's identity cannot originate from a registered secret; it must be attested by the execution platform, ephemeral, automatically rotated, and presented over an authenticated channel. Two independent checks then coexist: proof of possession binds the credential to a key, while channel authentication binds the connection to a process. Neither of them asserts that key and process belong to the same executor. A second attested process that obtains the credential and the key, both data at rest, presents valid proof over its own attested channel, and both checks pass, which amounts to what we call cross-plane impersonation. We argue for converting this apparent redundancy into a cross-plane binding: issuance records in the credential the attested identity of the recipient, verified at that moment, and the service compares it with the channel identity on every request. Binding the attested key itself is an alternative, at the cost of requiring a validity shorter than the rotation cycle. The attested identity belongs to the infrastructure plane (which process executes); the delegation asserted at issuance belongs to the organizational plane (on whose behalf, with which rights). The binding resolves cross-plane impersonation in the local case, a single domain whose issuer verifies the attestation. Evaluating both planes together is the role of the next layer; reconciling them in general remains open (Section 5).

The third layer decides whether this request, from this agent, on behalf of this human and in this context, should be served. In the decision document, the service assembles the identity of the delegator and their permissions, the attested identity of the executor, the intent declared in structured form, and the operational context, and queries a decoupled policy engine through a standardized interface. The decision, which may carry obligations such as masking attributes, is enforced by the service and recorded in a tamper-evident audit trail.

3. Authorization Flow and Policy Principles

The flow runs through the steps of reception; verification of the credential and of the proof of possession; attestation of the executor, including the correspondence between the key confirmed by the credential and the channel identity; assembly of the decision document; evaluation by the policy engine; provisioning with minimal scopes and validity, or denial that distinguishes an authentication failure from an authorization denial; and auditing.

This yields fast denial at the edge and reproducible decisions, since the decision document is versioned along with the policies. Four principles, composed by conjunction under deny-by-default, govern the flow: **temporal conformance**, which restricts windows and validity periods; **least privilege by intersection**, by which the effective permission is the intersection of the delegator's rights and the executor's attested capabilities, never their union; **intent coherence**, which compares the declared intent with the requested operation; and **personal data governance**, which minimizes and masks attributes whenever the decision so requires.

4. Discussion and Related Work

The arrangement covers the central threats. Credential theft and replay run up against proof of possession; executor impersonation, against attestation; cross-plane impersonation, against the cross-plane binding; delegation escalation, against the intersection; intent deviation, against coherence; and misuse of personal data, against governance. Three limitations delimit its reach: the portability of agent identity across domains remains open [IETF WIMSE Working Group 2026]; the proposal presupposes a credential issuer able to express intent and delegation; and intent coherence verifies structural consistency, not the agent's internal plan [Syros et al. 2026]. The position can be instantiated with blocks that are standardized or being standardized, such as proof of possession of credentials [Fett et al. 2023, Campbell et al. 2020], explicit delegation [Jones et al. 2020], declaration of intent [Lodderstedt et al. 2023], attested workload identity [SPIFFE Project 2025], decision interface [OpenID Foundation 2026], and cross-domain identity [IETF WIMSE Working Group 2026]. None of them, however, prescribes on its own the composition or the cross-plane binding argued for here. The literature has established the problem [South et al. 2025a, South et al. 2025b, Tallam 2026], but without specifying, as an integrated whole, the separation between cryptographic pre-authorization, externalized decision, and cross-plane reconciliation.

5. Conclusion and Research Agenda

We argue that safely admitting agentic requests requires continuous verification on a per-request basis, materialized in three conceptual layers: cryptographic pre-authorization with proof of possession, attested identity bound to that proof, and externalized contextual decision. Together, they mark the shift from credential-centered authorization to an authorization centered on attested identity, delegation, intent, and context. The agenda has three fronts: a portable representation of the agent's identity and capabilities beyond the local domain [IETF WIMSE Working Group 2026]; metrics and reproducible attack suites for agentic authorization [Tallam 2026]; and formalization of how the principles compose and of the guarantees this offers. A reference implementation is the next step.

Acknowledgements

Rodrigo acknowledges the support of FAPERGS (grant no.25/2551-0002572-9) and FAPESP (grants no. 2020/05183-0 and 2023/00816-2).

Referências

Campbell, B., Bradley, J., Sakimura, N., and Lodderstedt, T. (2020). Oauth 2.0 mutual-tls client authentication and certificate-bound access tokens. In *RFC 8705*. IETF.

Fett, D., Campbell, B., Bradley, J., Lodderstedt, T., Jones, M., and Waite, D. (2023). Oauth 2.0 demonstrating proof of possession (dpop). In *RFC 9449*. IETF.

IETF WIMSE Working Group (2026). Workload identity in a multi system environment (wimse) architecture. In *Internet-Draft draft-ietf-wimse-arch-08, Work in progress*. IETF.

Jones, M., Nadalin, A., Campbell, B., Bradley, J., and Mortimore, C. (2020). Oauth 2.0 token exchange. In *RFC 8693*. IETF.

Lodderstedt, T., Richer, J., and Campbell, B. (2023). Oauth 2.0 rich authorization requests. In *RFC 9396*. IETF.

OpenID Foundation (2026). Authorization api 1.0 — final specification. In *OpenID AuthZEN Working Group*. OpenID Foundation (Available at: https://openid.net/specs/authorization-api-1_0.html).

South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C. D., Greenwood, D., Chan, A., and Pentland, A. (2025a). Authenticated delegation and authorized ai agents. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025), Position Paper Track*. arXiv:2501.09674.

South, T., Nagabhushanaradhya, S., Dissanayaka, A., Cecchetti, S., Fletcher, G., et al. (2025b). Identity management for agentic ai: The new frontier of authorization, authentication, and security for an ai agent world. In *OpenID Foundation Whitepaper*. arXiv:2510.25819.

SPIFFE Project (2025). The spiffe identity and verifiable identity document (spiffe-id). In *Cloud Native Computing Foundation*. Available at: https://github.com/spiffe/spiffe/blob/main/standards/SPIFFE-ID.md.

Syros, G., Suri, A., Ginesin, J., Nita-Rotaru, C., and Oprea, A. (2026). Saga: A security architecture for governing ai agentic systems. In *Network and Distributed System Security Symposium (NDSS 2026)*. The Internet Society.

Tallam, K. (2026). Authorization propagation in multi-agent ai systems: Identity governance as infrastructure. In *arXiv preprint*. arXiv:2605.05440.