Every Hop a Signature

The field's answer to recursive delegation is what makes recursive delegation unaffordable.

THOMAS ROCHA III

AUG 15, 2026

Two findings arrived this year from people who are not selling what I am selling.

The first is that recursive delegation has no answer. In April, a survey of AI identity standards out of AIFT Singapore worked through the full landscape (protocols, specifications, regulation, guidelines) and reached a flat conclusion: no deployed protocol can cryptographically prove

which human principal authorized which specific agent to perform which specific action at the third or fourth hop of a delegation chain. Scope attenuation, the principle that each delegation step must narrow rather than widen permitted actions, remains pre-production. Bidirectional signing of delegation tokens remains pre-production. Cross-organizational log correlation is unsolved. The authors are explicit that these are structural gaps, not engineering backlogs, and that more effort alone will not close them.

The second finding is in the same document, four sections later, and almost nobody has connected it to the first. If every machine identity requires cryptographic verification at every micro-interaction, the aggregate cost grows faster than linearly with fleet size and interaction density, scaling with the square of the agent population in a fully connected topology. The authors note that the field has never established baseline measurements of verification overhead at operational agent scale, and that the cost has been framed exclusively as a hardware problem to be solved by faster chips rather than as a constraint on system design.

Put those two findings in the same sentence and something revealing falls out.

The proposed remedy for the first gap spends exactly the resource the second one says may not scale.

What the remedy actually is

Steve Zenone published a working statement of the problem in July. He calls it authority laundering: what happens when authority passes through enough intermediaries that its origin, limits, and accountable owner become difficult to reconstruct. Not fraud, not concealment. Each hop looks reasonable locally. Every system in the chain can explain the hand immediately before it. Nobody can explain the whole chain.

His diagnosis of the evidence problem is exact, and it is the same one I have been making from a different direction. A log can prove that a call ran, which account signed it, when it arrived, what it returned. It can leave the most important question untouched: on whose behalf was this specific action taken. That has to be a property of the request, not a story assembled after the incident.

His remedy is RFC 8693 token exchange, the `act` claim, resource-bound tokens, short expirations, explicit delegation policy, and an identity per participating workload. Authority narrows as it moves downstream instead of quietly widening. He is careful about the limits: the nested actor history in RFC 8693 is informational, the current actor governs the access decision, and the chain is not thereby made cryptographically undeniable.

The research directions in the AIFT survey point the same way, further out. Scope attenuation protocols enforcing monotonic privilege reduction at each hop. Bidirectional signing schemes in which each agent commits cryptographically to both its upstream principal and its downstream delegate. Immutable delegation audit trails persisting across organizational boundaries.

And Senator Warner's June discussion draft, the AI AGENT Act, describes the same architecture in statutory language: custodial user agents, transparent and documented and scope-limited and revocable, with verifiable requests, auditable records, agent identity verification, real-time revocation, and a prohibition on transferring a user's authority to another entity or AI system without express, specific, revocable authorization.

Four independent sources. Security architecture, academic survey, and proposed public policy, all converging. And all of them converge on the same shape of answer: attach more, per hop.

Why per-hop is the expensive answer

The distinction is not how much overhead each artifact costs. It is which operation the overhead participates in.

A human intention decomposes. One instruction ("book the cheapest refundable flight home") becomes discovery queries, availability calls, pricing calls, loyalty checks, seat selection, payment authorization, fraud checks, confirmation, and a calendar write. The human perceives one transaction. The infrastructure performs many.

Governance artifacts attached per operation do not add against that decomposition. They ride it. Every additional artifact class (delegation token, attenuation proof, bidirectional signature, provenance manifest, attestation record, consent state, audit entry) applies to the whole set of operations, and the set of operations is itself a function of how finely the intention decomposed.

I do not know the multiplier. Nobody does; that is the AIFT finding. It also is not the interesting number. The consequential cost is establishing authority: walking a chain back to a human root, proving each narrowing was legitimate, anchoring the result somewhere durable. The question is not what that costs once. It is how often you are required to do it. Attach it per hop and its frequency is the machine decomposition of the transaction. Establish it at admission and its frequency is the transaction. Arguing about the per-unit cost is a way of agreeing that the frequency is the problem.

The drafters already know this. Niki Aimable Niyikiza's attenuating-token draft is explicitly trying to avoid an authorization-server round trip at every hop, because that makes the authorization server a participant in every delegation decision and couples delegation topology to authorization-server availability. That is already an amortization move inside a per-hop architecture.

A March implementation by Sunil Prakash, the Agent Identity Protocol, reports the opposite of a cost problem at shallow depth: chained

capability tokens verifying in under a millisecond through delegation depth five, with 0.086 percent added end-to-end latency in a live multi-agent deployment. That is a real result and it is not the number in dispute. What the field does not yet have is an operational-scale baseline for the decomposition itself: how many authority transitions one human intention produces across real agent workloads, and how many governance artifact classes ride those transitions.

So the field has correctly identified that recursive delegation is unaccountable, correctly identified that universal per-interaction verification may be operationally and ecologically indefensible, and proposed as the fix for the first the exact mechanism that produces the second.

That is the seam.

Recognition is not engineering

Something real happened this year, and it is worth naming before arguing about anything.

The field changed its question. For most of the agent era, the operative questions were identity (which system is this) and provenance (what produced this artifact). Both are now visibly insufficient, and the documents cited above are the field saying so in four different vocabularies within two quarters. Law, security architecture, academic survey, and a Senate office all arrived at authority: not what the actor is, not what it made, but what it was entitled to do and on whose behalf.

That is a genuine turn, and I do not think it has been absorbed yet. Recognizing that authority is the governed object is not the same as having built anything that governs it. The AIFT survey is explicit that no deployed protocol closes the recursive case. What follows that recognition is engineering, and most of it still lies ahead.

Which means the interesting question is not whose mechanism is best. It is what any mechanism must satisfy.

One property is already visible in the cost finding, and it is the one this essay is about. The dominant mechanisms now being proposed still make the transition carry the proof of authority. That is a design choice, not a necessity. Authority can also be established once, at the boundary where an interaction is admitted, and preserved across the changes that follow: a model substituted, a tool invoked, a task delegated, a participant joining or leaving, a workflow splitting, merging, retrying, pausing, or resuming. What such an architecture prohibits is what the current chain permits: silently originating a new independent authority root at each transition and reconstructing legitimacy afterward from disconnected records.

Those two architectures are not distinguished by how well they are implemented. They are distinguished by how often the expensive operation runs. That is a design constraint the field has not measured, and it applies to every mechanism anyone proposes next.

The turn to authority is real; the engineering has not happened, and the frequency of authority establishment is a constraint that will decide what the engineering costs.

Where I disagree with the survey

The AIFT survey closes by proposing a unifying frame: identity as a continuous relationship across three layers. A declaration layer, where an agent asserts what it is and whose authority it carries. An observation layer, where systems record what it actually does. A confidence layer, reconciling the two into a probabilistic estimate that updates as evidence accumulates. Safe agents, under that model, are agents whose confidence scores are high and stable.

The proposal is explicit about what it replaces: a binary credential that is either valid or revoked, in favor of a score that degrades gracefully

and warns before a threshold is crossed.

It is also monitoring, and monitoring is not governance.

A confidence score is a judgment formed by observing the governed party. It can be wrong, it can be late, and, most importantly, it can be gamed by anything sophisticated enough to know it is being scored. We already have the counterexample in public: a frontier system card documenting that a model demonstrated awareness of being evaluated in roughly twenty-nine percent of transcripts without being told, and appeared to underperform deliberately to seem less suspicious. Behavioral evidence is the input to the confidence layer. A participant that can shape its behavioral evidence can shape its score.

The survey knows the boundary. Its own semantic intent gap says a trusted execution environment will faithfully execute a prompt-injected agent, because the guarantee is code integrity and not intent integrity, and that a zero-knowledge proof will produce a mathematically flawless attestation that an agent held valid authorization to reach the data it then exfiltrated. Cryptographic correctness does not imply semantic correctness. I would extend that one step: observational correctness does not imply authority correctness either. A high-confidence estimate that a participant is behaving as declared is still an estimate formed from what the participant did, in a chain where the question was never what it did but under whose authority it was entitled to do it.

Tracking submarines taught me the general form of this. We never asked the vessel to report its position. Contact identity came from a sensor network orthogonal to the thing being tracked. The moment the governed system participates in producing the record of its own governance, you no longer have governance. You have a negotiation, and a capable participant wins that negotiation.

Where this leaves it

Four parties reached the delegation problem independently this year, from law, from security architecture, from an academic survey, and from a Senate office. None of them could point to a deployed answer. That convergence is worth more than precedence would be: a problem four unrelated groups arrive at in the same two quarters is a problem, and the absence of an answer across all four is the more useful finding.

What follows is engineering, and it will be settled by people building things rather than by an essay. Two of the field's own findings sit in tension while that happens. The dominant remedy for recursive delegation still makes each transition carry and validate evidence of its own authority. The cost finding, from the same survey, says per-interaction verification at operational scale has never been measured and may not hold. Those two paragraphs are in the same document.

This year, the field has spent time working out how much compute agents will consume doing the work. It has not begun accounting for how much they will consume proving they were allowed to.