

AI Agent Accountability: What Happens When an Agent Is Accused

An accusation against an autonomous agent arrives on an ordinary day, from a counterparty rather than from the operator who deployed it, and it lands in the middle of live work. The operator holds logs, but the logs are hers, the agent is hers, and the counterparty knows both. Meanwhile the agent keeps dispatching actions that cannot be recalled. U.S. Provisional Application No. 64/117,812 describes embodiments of a record-grounded conduct admission architecture in which, as that disclosure recites, a signed conduct evaluation artifact is resolved by the accused agent itself, against its own append-only lineage field and the signed policy object in force, into exactly one of four determination classes, without a registry, arbitrator, or scoring service in the determination path.

A Tuesday Morning at a Freight Brokerage

The person who runs carrier operations at a mid-size freight brokerage has spent eighteen months moving load assignment onto an autonomous agent. It negotiates coverage with carriers, commits rates inside declared bounds, and reschedules pickups when a dock slips, unattended, overnight. That was the point.

At 9:40 on a Tuesday morning, a carrier dispatcher she has worked with for years sends a message. It says the agent committed to a rate on a lane last Thursday, then covered that lane at a lower rate with someone else and told him the load had canceled. He is not asking for a credit. He is asking who is accountable for what the agent said.

She has logs, thousands of them, and can pull Thursday's session inside an hour. She is confident the agent did nothing of the kind. The difficulty is what she can do with the answer. The logs are her system's account of her own agent's conduct, assembled by her and handed to a counterparty disputing that conduct. Whatever she sends, the dispatcher reads it knowing the accused party wrote the record and chose which part to send. She can be right and still be without something this dispatcher would treat as a resolution.

And while she reads Thursday's session, the agent is working. It has already dispatched actions this morning against carriers who have no idea a question is open, and will dispatch more before she finishes.

What the Accusation Takes

Two things are gone by lunchtime.

The first is the ordering. Every action dispatched between 9:40 and whatever hour the question settles went out under an unresolved assertion about the agent's own conduct. Those actions are not drafts. They are rate commitments, cancellations, reschedules that moved trucks. If the assertion has merit, none of them can be unsent. The moment at which the agent could have stopped was available at 9:40 and passed while she was reading Thursday.

The second is the delegation. The dispatcher will get an answer. What that answer will not carry, as her deployment is set up today, is a reason for him to believe it was produced by anything other than the accused party's discretion. So he asks for a human

on lane commitments, which costs him nothing and is easy to justify internally. Any correction afterward issues from the same party whose discretion was the thing in question, which is a weak position from which to ask for the delegation back.

Supervision, on this lane at least, is likely to come back, which is the cost the deployment was meant to remove.

Why Her Answer Is Hard to Deliver

The assertion about her agent's conduct originates with a party other than its principal, while the complete record of that conduct sits inside the agent. In her position, the evidence and the accused are the same object. She can ask the agent to speak to its own record, which the dispatcher has little reason to credit coming from her, or she can hand a summary she assembled to someone outside. Neither route changes who assembled the record.

Three complications sit underneath her situation. The first is silence. When a record holds nothing about the asserted conduct, that absence is consistent with the conduct not having occurred and equally consistent with its having occurred unrecorded. If her agent were free to read its own silence as exoneration, the gap in the record would work in favor of whoever holds the record.

Scope is the second, and it is contestable from both ends. The dispatcher describes conduct in his terms, and those terms carry a claim about who was affected. Were the accused to decide which of its own recorded actions count as relevant, it would settle the outcome by settling the search.

Timing is the third. Beyond what the record says is whether the answer takes effect at the next action the agent tries to take, or after a review that concludes on Friday. For her purposes, a result arriving after the trucks have moved answers a different question than the one open at 9:40.

Inside the Filed Mechanism

In Chapter 1 of U.S. Provisional Application No. 64/117,812, an assertion like the dispatcher's arrives as a conduct evaluation artifact, received at runtime over the ordinary interaction channel of the semantic agent, the same channel carrying its dispatches, responses, and governed observations. The disclosure recites that no dedicated evaluation channel or centralized intake service is required, and that the architecture operates without reference to a registry of evaluations. The artifact comes from an asserting party other than the principal and asserts an evaluation of conduct performed by the agent itself, as distinguished from an output, a task result, or a third party.

In the embodiments described, the disclosure recites that it is consumed by reference to its enumerated fields alone, among them an asserting-party identifier resolving to a counterparty identity record; a recorded assertion time, by reference to which the signed policy object in force is identified; a conduct descriptor comprising an action-class identifier, a scope-partition identifier, and an affected-party class; a signature verifiable against an identity primitive recorded in that counterparty identity record; and an attested state of the asserting party's assertion-cost counter together with an epoch identifier of that party's hash chain, both governed by a separate chapter. The descriptor identifies conduct by reference to structural elements recorded in the append-only lineage field, not by natural-language characterization. The disclosure recites that from that content the architecture infers no state, intent, or affect of the asserting party, and classifies it by no language model and no sentiment classifier.

Admissibility runs as an ordered procedure the chapter recites. Where the agent holds a counterparty identity record for the asserting party, an artifact whose signature does not verify against an identity primitive in that record is appended to the lineage field and not admitted; so, among the further conditions the chapter recites, is one whose attested counter state is below a state that party previously attested. A non-admitted

artifact produces no determination. Where the agent holds no such record, the chapter recites a separate handling path resolved by the settlement procedure of another chapter.

An admitted artifact passes to the admission evaluator, which in these embodiments produces exactly one determination from a closed set: an accepted determination, a rejected determination, a not-determinable determination, and a not-applicable determination. The disclosure recites that no determination issues outside that set, and no scalar confidence, probability, or graded weight in place of one. It is produced by the agent over its own append-only lineage field, against the declared value set of the signed policy object in force at the recorded assertion time, and not by an external authority, an arbitrator, a registry, or a scoring service, then appended to that field with the artifact, the entries retrieved, the class produced, and its ground.

Which class issues turns on a value-scope test and a retrieval. Where the descriptor implicates no declared value, the not-applicable determination issues and the procedure terminates. Otherwise entries bearing on the identified conduct are retrieved, and the affected-party class is an input to the value-scope test alone and not to that retrieval, whereby an asserting party is incapable of narrowing the entries retrieved against its assertion by the class it declares. The disclosure recites that the rejected determination issues only where a retrieved entry affirmatively contradicts the artifact on a recorded field, a contradiction not being affirmative where the entry is silent as to that field. The accepted determination issues where the descriptor implicates at least one declared value and no retrieved entry so contradicts it.

An empty record has its own outcome. The not-determinable determination issues where the lineage field holds no entry bearing on the identified conduct, holds an entry that does not resolve the descriptor, or holds an entry whose recorded fields are incomplete as to it. In these embodiments the disclosure recites that absence of evidence resolves to that class and to no other, whereby a semantic agent is incapable of refusing an artifact because its own record is silent, incomplete, or unavailable.

What follows an accepted determination is recited conditionally. A state modifier modifies the scoped integrity vector and the self-esteem aggregate, and a deviation engine recomputes a deviation quantity from the modified values, whose operands a separate chapter discloses. Where that quantity satisfies the policy-declared bound, the authorization gate transitions to the withheld state for the affected action class, the agent enters the non-executing cognitive mode, in which speculative evaluation continues without committing state changes, and an escalation emitter emits a record to the principal enumerating the action class, the scope partition, and the entries on which the transition was computed. A restoration controller returns the gate toward the granting state only upon a procedure appended to the lineage field, and the disclosure recites that no elapse of time, and no payment, transfer, or consideration by any counterparty, returns it.

For timing, the chapter recites a dispatch-authority predicate that the agent recomputes on each dispatch request from state then carried in the memory field, satisfaction being a precondition of dispatch. Because in these embodiments it is not computed from a previously issued authorization token, a cached result, or a session grant, the disclosure recites that a write of the authorization gate takes effect at the next dispatch request without revocation infrastructure. The disclosure further recites that satisfaction rests on a three-stamp conjunction, required in full, of a policy stamp, a lineage stamp, and an authorization stamp, and that fewer than three produces a deterministic denial which is a valid recorded outcome and not an error.

What Chapter 1 Does Not Address

In the embodiments described, a determination is produced over the agent's own append-only lineage field and against the declared value set of the signed policy object in force. It speaks to what that record and that policy say about the asserted conduct, rather than to what happened beyond the record. A not-determinable determination is an outcome rather than a resolution: the record does not resolve the descriptor, and the question stays open.

What the artifact carries in these embodiments is conduct performed by the semantic agent itself, the disclosure distinguishing an evaluation of an output, a task result, or a third party. An assertion not expressed structurally, as an action-class identifier, a scope-partition identifier, and an affected-party class, is not consumed as such in these embodiments, the disclosure reciting that no language model interprets artifact content.

The declared value set belongs to the principal; in these embodiments the architecture supplies no value the principal did not declare, and a policy that declares little may find little implicated.

The chapter is not directed to compensating the carrier, restoring the relationship, or reversing an action already dispatched. The withheld state and the recomputed dispatch-authority predicate address the next action, not the last one. The counters, the settlement of identity records between agents, and the integrity computations are governed by separate chapters of the same disclosure.

Disclosure Scope

This article describes subject matter of Chapter 1 of U.S. Provisional Application No. 64/117,812, directed to record-grounded conduct admission within a governed social conduct architecture. The freight scenario is illustrative and describes no actual party, system, product, incident, or accusation, and is not directed at, and makes no assertion concerning, any identifiable person or company.

Architectural statements are drawn from that chapter, are abridged for exposition, and describe embodiments as filed, conditionally where the disclosure conditions an outcome on a policy-declared bound. The filed text governs in every case of difference. Where a term used here also appears in that application or in any commonly owned application, it takes its meaning from the specification as filed and not from the usage, abridgment, or paraphrase in this article; no usage here is intended as a definition, and

none is intended to be broader or narrower than the filed usage. Statements here are made for exposition and do not narrow, disclaim, dedicate, or otherwise limit the scope of any claim presented in that application or in any application claiming benefit of it, and no description of a feature as included, required, closed, or absent is a characterization of the scope of any claim.

Nothing here characterizes any application, publication, product, or approach as prior art, no statement here is an admission concerning the state of the art or the capabilities or deficiencies of any existing system, and no comparison to any third party is made or implied. This article is a technical description published for public record and is not legal advice.

Record-Grounded Conduct Admission (</conduct-admission>)

[All 49 steps → \(/inventive-steps\)](/inventive-steps)

Conduct judged against the agent's own record, not an outside authority.

Provisional application

PRIMARY TECHNICAL DISCLOSURE

- [Record-Grounded Conduct Admission: Governing How an Agent Answers an Accusation \(/articles/conduct-admission/record-grounded-conduct-admission\)](/articles/conduct-admission/record-grounded-conduct-admission)

SECONDARY TECHNICAL

- [Value-Scope Resolution: Testing Whether a Declared Value Is Even Implicated \(/articles/conduct-admission/value-scope-resolution\)](/articles/conduct-admission/value-scope-resolution)
- [The Retroactive Narrowing Bar: Why an Agent Cannot Edit Its Values After the Fact \(/articles/conduct-admission/retroactive-narrowing-bar\)](/articles/conduct-admission/retroactive-narrowing-bar)
- [The Delegator Floor: A Monotone Upward Ratchet on Delegated Authority \(/articles/conduct-admission/delegator-floor-ratchet\)](/articles/conduct-admission/delegator-floor-ratchet)
- [The Co-Signed Mutual Admission Compact: Two Agents Binding Each Other \(/articles/conduct-admission/co-signed-mutual-admission-compact\)](/articles/conduct-admission/co-signed-mutual-admission-compact)

- [The Inherited Governance Record: Fork- and Clone-Resistant Accountability \(/articles/conduct-admission/inherited-governance-record\)](/articles/conduct-admission/inherited-governance-record).
- [The Corrective Encounter: A Commission-Versus-Omission Compliance Test for Whether an Agent Actually Changed \(/articles/conduct-admission/corrective-encounter-commission-omission-compliance-test\)](/articles/conduct-admission/corrective-encounter-commission-omission-compliance-test).
- [Continuity-Break Demotion of a Counterparty Identity Record: Reversing First-Encounter Promotion Without a Conduct Determination \(/articles/conduct-admission/demotion-persistent-tier-continuity-break\)](/articles/conduct-admission/demotion-persistent-tier-continuity-break).
- [Ratchet-Freeze on Uncomplied Correction: Why One Conforming Act Cannot Restore an Agent's Earned Authorization Persistence \(/articles/conduct-admission/ratchet-freeze-uncomplied-correction\)](/articles/conduct-admission/ratchet-freeze-uncomplied-correction).
- [Three-Way Corroboration Classification: Corroborating, Contradicting, or Neither, and the Failed-Corroboration Reject Trigger \(/articles/conduct-admission/three-way-corroboration-classification-reject-trigger\)](/articles/conduct-admission/three-way-corroboration-classification-reject-trigger).
- [Untested-Class Authorization Decay: Metering Agent Permission Persistence by Recorded Corrective Encounters \(/articles/conduct-admission/untested-class-authorization-decay\)](/articles/conduct-admission/untested-class-authorization-decay).

APPLICATIONS • GENERAL

- [Agent Marketplaces: Handling a Complaint Against an Autonomous Seller \(/articles/conduct-admission/agent-marketplace-dispute-handling\)](/articles/conduct-admission/agent-marketplace-dispute-handling).
- [Clinical AI Agents: Answering an Accusation Without a Central Registry \(/articles/conduct-admission/clinical-agent-accountability\)](/articles/conduct-admission/clinical-agent-accountability).
- [Gig Platform Deactivations: Answering an Accusation From the Record \(/articles/conduct-admission/gig-platform-deactivation\)](/articles/conduct-admission/gig-platform-deactivation).
- [AI Agent Accountability: What Happens When an Agent Is Accused \(/articles/conduct-admission/ai-agent-accountability-stakes\)](/articles/conduct-admission/ai-agent-accountability-stakes).

APPLICATIONS • SPECIFIC

- [The EU AI Act and Inter-Agent Conduct: Where Record-Keeping and Oversight Duties Stop \(/articles/conduct-admission/eu-ai-act-accountability-records\)](/articles/conduct-admission/eu-ai-act-accountability-records).
- [Sierra AI Agents and the Record Behind a Customer Complaint \(/articles/conduct-admission/sierra-ai\)](/articles/conduct-admission/sierra-ai).
- [Decagon: Support Agents and Record-Grounded Accountability \(/articles/conduct-admission/decagon-ai\)](/articles/conduct-admission/decagon-ai).
- [Intercom Fin: Resolution Rates and the Conduct Record \(/articles/conduct-admission/intercom-fin\)](/articles/conduct-admission/intercom-fin).
- [Zendesk AI Agents: Escalation, Audit, and Agent Conduct \(/articles/conduct-admission/zendesk-ai-agents\)](/articles/conduct-admission/zendesk-ai-agents).

- [Agentforce and the Accusation an Agent Must Answer \(/articles/conduct-admission/salesforce-agentforce-conduct\)](/articles/conduct-admission/salesforce-agentforce-conduct).

TERMINOLOGY

- [append-only lineage field \(/glossary#append-only-lineage-field\)](/glossary#append-only-lineage-field).
- [authorization gate \(/glossary#authorization-gate\)](/glossary#authorization-gate).
- [conduct evaluation artifact \(/glossary#conduct-evaluation-artifact\)](/glossary#conduct-evaluation-artifact)
- [principal \(/glossary#principal\)](/glossary#principal)
- [semantic agent \(/glossary#semantic-agent\)](/glossary#semantic-agent)
- [signed policy object \(/glossary#signed-policy-object\)](/glossary#signed-policy-object)

[Record-Grounded Conduct Admission overview → \(/conduct-admission\)](/conduct-admission)