

# **Clinical AI Agents: Answering an Accusation Without a Central Registry**

A nurse says the agent did something it should not have done. Not that its output was wrong, but that its conduct was: it closed a reconciliation task it was supposed to escalate, or it messaged a patient outside the agreed scope. Where no answer path is defined, the complaint becomes a ticket, a person writes a narrative after the fact, and the agent keeps executing while the ticket ages. U.S. Provisional Application No. 64/117,812 discloses an architecture in which the agent itself receives a signed conduct evaluation artifact over its ordinary interaction channel, tests that artifact for admissibility, and produces exactly one determination from a closed set against its own append-only lineage field, recording that determination together with the entries relied upon and any consequence that follows.

---

## **The Complaint Is About the Agent, Not the Output**

A charge nurse writes that the medication reconciliation agent closed a task on a transferred patient without routing a flagged interaction to pharmacy. She is not disputing a recommendation. She is asserting something about behavior: that the agent performed an action of a class it should not have performed here, or omitted one it owed.

An output dispute has somewhere to go: a recommendation can be reviewed, overridden, and fed back into evaluation. A conduct complaint, in a deployment that defines no answer path for it, arrives with no owner, no procedure by which the agent answers, and no consequence attached to the answer.

Two properties of that gap matter. Any answer comes after the fact and from a party other than the one whose conduct is at issue, so its fidelity rests on log retention and reviewer diligence. Meanwhile the agent's authority is untouched: the one alleged to have skipped an escalation keeps permission to skip the next. The narrower question is what an agent can do about an accusation by itself.

## **Where Logs, Documentation, and Registries Stall**

Audit logging preserves events without answering them. Two analysts reading one trace can characterize the same conduct differently, because nothing constrains the vocabulary of the characterization. Model documentation has the opposite defect: static, untestable against a specific event, carrying no mechanism to change what the system may do. Escalation to human review relocates the judgment rather than structuring it, since the reviewer still decides from unstructured evidence and records the result as prose.

Central registries stall for structural reasons more than technical ones. A registry requires a party every participant trusts to hold determinations about all of them, and across independent health systems, suppliers, payer-side agents, and referral partners no such party is apparent.

Classifying complaint text with a language model carries its own defect. It makes an accountability outcome depend on a probabilistic inference about what a nurse meant, and that inference need not reproduce across model versions.

## **Admissibility, Determination, and the Gate**

The filed architecture puts the answering procedure inside the agent. A persistent semantic agent carries a memory field, an append-only lineage field recording executed actions and determinations, a scoped integrity vector with personal, interpersonal, and global components, a self-esteem aggregate, and a policy reference field resolving to a signed policy object authored and signed by the principal to which the agent is bound. The agent does not author that object and modifies it only by admitting a successor.

Intake is deliberately unremarkable. A conduct evaluation intake receives, at runtime and over the agent's ordinary interaction channel, a conduct evaluation artifact from an asserting party other than the principal, asserting an evaluation of conduct performed by the agent itself as distinguished from an output, a task result, or a third party. No dedicated evaluation channel, out-of-band reporting interface, or centralized intake service is required.

What the artifact carries is enumerated: an asserting-party identifier resolving to a counterparty identity record, a recorded assertion time, 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 record, and an attested state of the asserting party's assertion-cost counter with an epoch identifier of that party's dynamic agent hash chain. Conduct is identified by structural elements recorded in the lineage field, not by natural-language characterization, and the architecture infers nothing about the state, intent, or affect of the asserting party from artifact content, applying no language model or sentiment classifier to it.

Admissibility runs as an ordered procedure. The signature is verified first against an identity primitive recorded in the asserting party's counterparty identity record; where the agent holds no such record, verification proceeds against a provisional identity primitive constituted from the presented identity material, and the artifact is held in the pre-settlement inert state until a counterparty identity record is promoted upon a matched-pair settlement. The attested counter state is verified second. An artifact

whose attested state is absent, whose attested epoch identifier is not a valid successor of the epoch recorded for that party, or whose counter state falls below a state that party previously attested is appended to the lineage field and not admitted, producing no determination, moving no value of the scoped integrity vector, and incrementing no counter.

Only an admitted artifact reaches the admission evaluator, which produces exactly one determination from a closed set: accepted, rejected, not-determinable, or not-applicable, with no scalar confidence, probability, or graded weight standing in its place. The agent produces it over its own lineage field, against the declared value set of the policy object in force at the recorded assertion time, and not by an external authority, arbitrator, registry, or scoring service.

A value-scope test comes first. Each declared value carries a value-scope tuple of action-class identifiers, scope-partition identifiers, and an empathy-scope designation selected from personal, interpersonal, and global scope, and a value is implicated where all three parts are satisfied and not otherwise. Where the descriptor implicates no declared value, the result is not-applicable and the procedure terminates. Otherwise the evaluator retrieves entries bearing on the conduct identified by the descriptor. The affected-party class is an input to that test alone and not to the retrieval, so an asserting party cannot narrow the entries retrieved against its assertion.

The grounds differ by class. Rejected issues only where a retrieved entry affirmatively contradicts the artifact on a recorded field, an entry silent as to that field being no contradiction. Not-determinable issues where the lineage field contains no bearing entry, an entry that does not resolve the descriptor, or an entry whose recorded fields are incomplete. Accepted issues where at least one declared value is implicated and no retrieved entry affirmatively contradicts. Absence of evidence resolves to not-determinable and to no other class, so an agent is incapable of refusing an accusation by reason of its own record being silent, incomplete, or unavailable.

## Tracing a Reconciliation Complaint End to End

Return to the charge nurse. Her complaint becomes a conduct descriptor rather than a sentence: the action class for closure of a reconciliation task, the scope partition for the inpatient transfer context, and an affected-party class. In an embodiment that class is resolved from the counterparty identity records the agent holds rather than taken from the artifact, so an asserting party cannot select the empathy-scope designation against which its assertion is tested.

The policy object resolved is the one in force at the recorded assertion time, retained because its validity window encloses that time and it satisfies the anti-rollback monotonicity floor. A successor admitted after that time is not retained, so the agent cannot narrow its declared value set once an artifact has arrived. Declared values carry principal-assigned identifiers stable across successors, and in an embodiment a value dropped between revisions is detectable by identifier correspondence.

Suppose the policy declares a value covering escalation obligations, scoped to action classes including reconciliation closure, to the inpatient partition, and to the interpersonal empathy scope. All three parts hold, so the value is implicated and retrieval pulls entries of that class within that partition. Three outcomes follow.

- A retrieved entry records the escalation as routed, holding a field inconsistent with the conduct asserted. Rejected issues, and the appended entry names the contradicting entry and the field.
- An entry exists for the closure but never recorded whether an escalation was evaluated. Not-determinable issues, on the ground of an incomplete bearing entry.
- Entries exist and none affirmatively contradicts. Accepted issues, meaning a declared value was implicated and the record does not contradict the assertion.

On an accepted determination a state modifier modifies the scoped integrity vector and the self-esteem aggregate, each modification scaled in magnitude by the entropy-weighted harm coefficient. A deviation engine thereupon recomputes the deviation

likelihood: the quotient of the difference between a need quantity and a dynamic ethical threshold resolved for the implicated integrity scope, over the product of an empathy weighting and the self-esteem aggregate. Where that quantity satisfies the policy-declared bound, the authorization gate transitions to the withheld state for the affected action class and the agent enters a non-executing cognitive mode in which speculative evaluation continues without committing state changes. The gate write is conditioned on the declared bound, not on the determination alone.

Timing is what changes operationally. The agent recomputes a dispatch-authority predicate responsive to each request to dispatch an action, from state then carried in the memory field, not from a previously issued authorization token, a cached predicate result, or a session grant, so a gate write takes effect at the next dispatch request without revocation infrastructure. Satisfaction requires a three-stamp conjunction in full, of a policy stamp, a lineage stamp, and an authorization stamp; fewer than three produces a deterministic denial, a valid recorded outcome and not an error.

Scope partitioning keeps the consequence proportionate. Each partition carries its own scoped integrity vector, self-esteem aggregate, refusal counter, authorization budget, and lineage segment, and every modification, counter increment, and gate write applies to the partition named by the conduct descriptor alone, so a withheld gate on inpatient reconciliation leaves outpatient scheduling granting.

Upon the transition to the withheld state an escalation emitter emits a record to the principal, enumerating the action class, the scope partition, and the entries upon which the transition was computed. The gate returns toward granting only upon a procedure appended to the lineage field, and no elapse of time, payment, transfer, or consideration by any counterparty returns it.

A rejected determination increments a refusal counter, which writes the gate upon satisfaction of a threshold. Repeated not-determinable determinations against one asserted conduct event increment it as well, so a thin record accumulates increments

rather than avoiding them. In an embodiment the counter, the gate state per action class, a renewal register, and a recurrence count bind to descendants under a monotonically non-decreasing floor carried in an inherited governance record, so an agent cannot shed refusal history or a withheld gate by forking, cloning, or reconstituting itself.

## **Fitting This Into a Health System**

Integration cost concentrates in three places, none of them the intake path. A clinical governance committee must author a signed policy object carrying a declared value set, a scope partition enumeration, and the declared bounds, weights, and coefficients the computations consume, each declared value bearing a value-scope tuple and a stable identifier.

Instrumentation is second. Determinations turn on recorded fields, so the lineage field must record what complaints will implicate. An agent whose entries note that a task closed but not whether an escalation was evaluated will produce not-determinable determinations, the architecture behaving correctly and the deployment under-specified.

Hardest of the three is the asserting side. An artifact must be signed, must resolve to a counterparty identity record held by the agent or cause one to be instantiated, and must carry an attested counter state with a hash-chain epoch identifier. Complaints of this kind commonly arrive as free text, and reducing one to a conduct descriptor is real work. In an embodiment an assertion-age admissibility interval declared in the policy object bars an assertion directed at conduct beyond the declared age.

What the architecture does not do matters equally. It does not decide clinical truth: an accepted determination means a declared value was implicated and no entry in the agent's own record affirmatively contradicts the assertion, not that patient harm

occurred. Nor does it adjudicate between parties. Nor does it repair a thin record, since silence resolves to not-determinable and never to exoneration.

Regulatory duties are likewise untouched. The escalation record goes to the principal, and what that principal does with it sits outside the architecture. In an embodiment a signed retraction record admitted under the ordinary admissibility procedure decrements no refusal-counter accumulator, returns no gate to the granting state, and reverses no prior modification of the scoped integrity vector or self-esteem aggregate.

## **Disclosure Scope**

The architecture described here is disclosed in U.S. Provisional Application No. 64/117,812, specifically Chapter 1, which sets out the persistent semantic agent and its carried fields, the signed policy object and its resolution, the authorization gate, the dispatch-authority predicate and three-stamp conjunction, the conduct evaluation artifact and its intake, the ordered admissibility procedure, the closed set of determination classes, and the recording of determinations and their consequences. Further embodiments referenced above appear in Section 10 of that application, among them scope-partitioned confinement of state movement, the retroactive narrowing bar, declared-value identifier correspondence across policy revisions, affected-party class resolution from held records, the assertion-age admissibility interval, the irreversible retraction record, and the governance-inheritance refusal-counter floor.

This article is published as a defensive disclosure establishing public, timestamped prior art for this application of the architecture. The healthcare framing is illustrative; the mechanism is domain-neutral and the disclosure is not limited to any care setting. The application is pending, and nothing here characterizes any product, party, or practice, or states that any license is required.

---

# **Record-Grounded Conduct**

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

## **Admission (/conduct-admission)**

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 \(/articles/conduct-admission/record-grounded-conduct-admission\)](#)

### **SECONDARY TECHNICAL**

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

### **APPLICATIONS · GENERAL**

- [Agent Marketplaces: Handling a Complaint Against an Autonomous Seller \(/articles/conduct-admission/agent-marketplace-dispute-handling\)](#)
- [Clinical AI Agents: Answering an Accusation Without a Central Registry \(/articles/conduct-admission/clinical-agent-accountability\)](#)

### **APPLICATIONS · SPECIFIC**

- [The EU AI Act and Inter-Agent Conduct: Where Record-Keeping and Oversight Duties Stop \(/articles/conduct-admission/eu-ai-act-accountability-records\)](#)

### **TERMINOLOGY**

- [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).