

When an Agent Mood Cannot Be Explained Afterward

On a Tuesday in the sixth week of her review, a compliance lead at a mid-size claims automation firm is asked one question she cannot answer: why did the agent handling a disputed file in March stop escalating? Her logs hold the outputs and the timestamps. They do not hold the disposition the agent was operating under in that hour, and in her deployment that disposition was overwritten by every update that followed. United States Patent Application 19/647,395 describes an architecture in which an agent's dispositional orientation is carried as a structured, policy-bounded field, in which every mutation to it is recorded in the agent's lineage, and in which the state at a queried timestamp may be reconstructed by replaying a deterministic update function over the recorded observations.

The Tuesday she cannot reconstruct

The compliance lead runs post-incident review for a claims automation firm with about four hundred employees. Her agents read intake documents, evaluate coverage against policy language, and either resolve a file or hand it to a human adjuster. Her review packets are usually dull, which is the point.

This one is not dull. A file from March went eleven days without a human touching it, and the agent that held it declined to hand it off four separate times before finally doing so on the twelfth day. Nobody disputes the outputs. The outputs are all in her logs, timestamped, with model versions and input documents attached. What the reviewing party on the other side of the table wants to know is narrower and harder: on the fourth of those four occasions, what was the agent's threshold for handing off, and why was it where it was.

She opens the record and finds what her deployment was configured to keep. She has the decision. She has the score that the decision was measured against. She does not have the setting that the score was compared to, because in her setup that setting is a live value in the running process, read at decision time and then updated, and the version of it that existed on the fourth occasion was replaced by the version that existed on the fifth. She can tell the reviewer what the agent did. She cannot tell the reviewer what the agent was like when it did it.

What her review loses, and why it stays lost

The thing she loses is not the incident. The incident is documented. What she loses is her firm's ability to say anything true about its own system in the one hour that matters, and that loss has a shape she recognizes immediately, because it is the shape of the findings she has spent six weeks writing about her own firm's systems.

Her review has two conclusions available to it now. She can say the agent behaved correctly, which she believes but cannot demonstrate, because demonstrating it would mean showing the disposition the agent was carrying and showing that the disposition sat inside the bounds her firm set. Or she can say the behavior is unexplained. The second conclusion is the honest one and it is the one that ends her ability to defend anything else the agent did that quarter, because the reviewer sitting across from her, having taken one unexplained escalation on faith, would have little reason to take the other eleven thousand on faith.

In her deployment it does not come back. This is the part she keeps returning to on the drive home. Had the March file been mishandled, she could have fixed the handling. Had the log been corrupted, she could have restored the log. But the state she needs was never a record in her deployment. It was a value that existed for a while and then did not, and she has no backup of a thing her system never wrote down. She can instrument the system tomorrow, and she will, and it will do nothing at all for March. Her firm will carry an unexplained interval in its own operating history for as long as anyone asks about that quarter.

Why the shape of her problem resists a late fix

She has considered simply logging more, and in her deployment that runs into three walls at once.

The first is volume. Were she to snapshot the full disposition of every agent at every decision point, she would be writing a high-frequency stream for a fleet that is mostly idle, for the benefit of the rare hour when someone asks. Her storage budget was set on the assumption that outcomes are what get kept.

The second is that the values she wants are, for her purposes, the most sensitive thing her agents hold. A stored history of exactly how cautious each agent was at each moment is a record she would then have to govern, retain, and eventually produce, and her privacy counsel has already told her what he thinks of accumulating a new class of behavioral state with no disclosure boundary.

The third is the one she finds hardest. Even with every value in hand, she is not sure she could show that the values were legitimate. In her setup there is no articulated account of what a permissible change to that state would have been, so a stored history would tell her the setting drifted without telling her whether the drift stayed inside

anything. The question put to her is not really what the number was. It is whether the number was governed. As configured today, her deployment holds no artifact that answers the second question even if she manufactures an answer to the first.

She also has a fourth problem she has not raised in the meeting yet, because she does not know how to bound it. Her agents delegate to one another. If the March agent's disposition was shaped by what a parent agent passed down at instantiation, then the interval she needs to explain does not begin with that agent at all, and nothing in her current setup lets her say where it does begin.

How the filed architecture keeps the record

United States Patent Application 19/647,395 describes, in accordance with an embodiment, an affective state field (200) introduced as a seventh structural field of the semantic agent schema, alongside the intent field, context block, memory field, policy reference field, mutation descriptor field, and lineage field. The disclosure is explicit that this field does not encode emotion in the phenomenological or subjective sense. It encodes a structured modulation vector that influences how the agent weighs alternatives, tolerates ambiguity, persists under partial failure, and escalates under constraint pressure.

In an embodiment, the field is organized as a modulation layer of named control fields, including an uncertainty sensitivity block (212), an ambiguity tolerance block (214), a novelty appetite block (216), a persistence under partial failure block (218), an escalation under time pressure block (220), an attention sensitivity block (222), and a cooperation disposition block (224). Each named control field is represented in an embodiment as a tuple comprising a current magnitude value within a defined range, a decay rate governing return toward baseline, a policy-defined ceiling and floor, and a timestamp recording the most recent update. The chapter describes these fields as modulating enumerated targets, among them promotion thresholds (226), search breadth (228), branch growth (230), escalation thresholds (232), decay rates (234), and

delegation routing (236). The escalation-under-time-pressure field is described as directly modulating escalation thresholds, with elevated uncertainty sensitivity and elevated risk sensitivity lowering them.

Updates are described as deterministic, such that the same agent state, the same environmental inputs, and the same policy configuration produce the same output. The update pipeline is depicted as a sequence: a structured observations block (238) into an update function block (240), then a policy bounds block (242), a decay curve block (244), a semantic hysteresis block (246), an entropy-governed stabilization block (248), and a lineage recording block (250). The policy-bounded update mechanism is described as verifying that an observation type is in the admissible trigger set for the relevant field, computing the raw update magnitude, clamping it to a rate limit, applying it, clamping the result to range bounds, and recording the complete transaction, including the observation, the raw update, the clamped update, the prior value, and the resulting value, in the agent's lineage.

From that combination the chapter describes affective state forensic reconstruction. Where a governance event requires post-hoc review of agent behavior, for example to determine the agent's modulation configuration at the time of a disputed decision, the disclosure describes reconstructing the state at any historical point by replaying the deterministic update function over the sequence of recorded observations from the lineage, without requiring persistent storage of moment-to-moment values. The lineage entries are described in an embodiment as abstract, recording the observation type, the update direction on each affected field, and the policy compliance status, rather than absolute field values or raw observations.

For delegation, the chapter describes an inheritance mask block (264) evaluated per named control field as inherited, excluded, or attenuated, a blending function block (266) producing a child affective state block (268), a return path block (270) feeding a parent update block (272), and a policy-specified maximum delegation depth beyond which inheritance does not occur. Blending operations are described as recorded in

both the parent's and the child's lineage, establishing a traceable affective provenance chain. Where volatility is detected, the chapter describes a volatility detector block (276) routing an agent into a quarantine state block (278) and a restricted mode block (280) with elevated promotion thresholds, suspended delegation authority, and elevated mutation acceptance thresholds, with release through a hysteretic recovery block (282) when the composite volatility metric falls below a recovery threshold.

The chapter is also explicit about separation. The affective state field is described as not an input to the governance gate, which evaluates execution admissibility on policy compliance, trust slope validation, and cryptographic provenance independently of the agent's affective state.

Where this would still leave her holding questions

The disclosed architecture would give her a governed answer to what the disposition was and whether its evolution stayed inside declared bounds. It would not, for her purposes, tell her whether the bounds her firm chose were the right ones. Setting a ceiling low enough that an agent under time pressure hands off within an hour rather than on the twelfth day remains a policy judgment she would have to make and defend, and the chapter describes those ceilings and floors as policy-defined rather than fixed by the architecture.

Nor would the abstraction level of the lineage records hand her the raw stream she originally imagined asking for. The chapter describes lineage entries in an embodiment as recording observation type, update direction, and compliance status rather than absolute field values, with reconstruction performed on demand by replay. That is a different artifact from a stored time series, and were her retention procedures written expecting one, she would have to rewrite them.

The reconstruction property described in the chapter rests on deterministic updates. The disclosure also contemplates alternative embodiments in which the update function incorporates bounded stochastic components, provided the stochastic contribution is policy-bounded and auditable through lineage recording. Were her deployment configured that way, she would need to understand the character of what she could reproduce before promising a reviewer exact replay.

The separation of concerns cuts against one hope she has been carrying. Because the field is described as not an input to the governance gate, a reconstruction would explain how her agent deliberated and would not by itself explain why a given action was admitted or denied. For the March file she would get the disposition. For the adjacent question of authority she would still be reading her governance record, a separate artifact with gaps of its own.

Her fourth problem narrows rather than disappears. Where inheritance is depth-limited by policy, an agent instantiated beyond that depth begins from its own baseline, so a provenance chain she traced backward would end at a defined boundary rather than at an original cause. That is more than she has today and less than the full causal story she went into the meeting wanting.

Disclosure Scope

This article describes subject matter disclosed in United States Patent Application 19/647,395. It is a technical description written for practitioners. Nothing in this article characterizes the scope of any claim, and nothing here is an admission regarding the state of the art. The scenario described is illustrative and does not depict any actual party, deployment, or event. Descriptions above refer to embodiments as disclosed, and where the disclosure conditions an outcome on a policy bound, threshold, or configuration, that condition is stated.

Affective State (</affective-state>)

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

Emotion as a computational primitive, not a simulation.

Chapter 2 (</patents/19-647395/chapters/affect>)

PRIMARY TECHNICAL DISCLOSURE

- [Affective State as a Deterministic Control Primitive for Semantic Agents \(/articles/affective-state-as-a-deterministic-control-primitive-for-semantic-agents\)](/articles/affective-state-as-a-deterministic-control-primitive-for-semantic-agents)

SECONDARY TECHNICAL

- [Affective State as the Seventh Structural Field \(/articles/affective-state/seventh-canonical-field\)](/articles/affective-state/seventh-canonical-field)
- [Named Control Field Modulation Architecture \(/articles/affective-state/named-control-fields\)](/articles/affective-state/named-control-fields)
- [Affect-Modulated Promotion Thresholds \(/articles/affective-state/promotion-thresholds\)](/articles/affective-state/promotion-thresholds)
- [Deterministic Affect Encoding and Update Mechanics \(/articles/affective-state/deterministic-encoding\)](/articles/affective-state/deterministic-encoding)
- [Emotional Decay Curves With Hysteresis \(/articles/affective-state/decay-curves\)](/articles/affective-state/decay-curves)
- [Entropy-Governed Valence Stabilization \(/articles/affective-state/valence-stabilization\)](/articles/affective-state/valence-stabilization)
- [Affective Inheritance in Delegation Chains \(/articles/affective-state/inheritance-chains\)](/articles/affective-state/inheritance-chains)
- [Emotional Quarantine and Volatility Management \(/articles/affective-state/emotional-quarantine\)](/articles/affective-state/emotional-quarantine)
- [Affect-Modulated Trust Slope Validation \(/articles/affective-state/trust-slope-modulation\)](/articles/affective-state/trust-slope-modulation)
- [Biological Signal-to-Affective Coupling \(/articles/affective-state/biological-coupling\)](/articles/affective-state/biological-coupling)
- [Affective Contagion in Multi-Agent Systems \(/articles/affective-state/affective-contagion\)](/articles/affective-state/affective-contagion)
- [Affect-Modulated Discovery Traversal \(/articles/affective-state/discovery-traversal\)](/articles/affective-state/discovery-traversal)
- [Affect-Governance Separation \(/articles/affective-state/governance-separation\)](/articles/affective-state/governance-separation)
- [Policy-Bounded Affective Updates \(/articles/affective-state/policy-bounded-updates\)](/articles/affective-state/policy-bounded-updates)
- [Affect as Cross-Primitive Input \(/articles/affective-state/cross-primitive-input\)](/articles/affective-state/cross-primitive-input)
- [Affect-Modulated Inference Integration \(/articles/affective-state/inference-integration\)](/articles/affective-state/inference-integration)
- [Substrate-Agnostic Affect Deployment \(/articles/affective-state/substrate-deployment\)](/articles/affective-state/substrate-deployment)
- [Pseudonymous Emotional Operation \(/articles/affective-state/pseudonymous-operation\)](/articles/affective-state/pseudonymous-operation)
- [Temporal Cognition Field \(/articles/affective-state/temporal-cognition\)](/articles/affective-state/temporal-cognition)

APPLICATIONS · GENERAL

- [How to Build a Companion AI That Keeps Emotional Consistency Across Sessions \(/articles/affective-state/companion-consistency\)](/articles/affective-state/companion-consistency)
- [Therapeutic AI Agents: Governed Emotional State for Mental Health Software Under Clinical Constraints \(/articles/affective-state/therapeutic-affect\)](/articles/affective-state/therapeutic-affect)
- [Persistent Affective State for Customer Service AI Agents: Beyond Per-Message Sentiment \(/articles/affective-state/customer-service-agents\)](/articles/affective-state/customer-service-agents)
- [AI Companion Agents for Elderly Care: Emotional Continuity and Mood Monitoring with Deterministic Affective State \(/articles/affective-state/elderly-care-companions\)](/articles/affective-state/elderly-care-companions)
- [Crisis Response AI Agents That Stay Calm: Governed Affective State for Emergency Operations \(/articles/affective-state/crisis-response-agents\)](/articles/affective-state/crisis-response-agents)
- [Auditable AI Negotiation Agents with Deterministic Affective State \(/articles/affective-state/negotiation-agents\)](/articles/affective-state/negotiation-agents)
- [Emotionally Adaptive AI Tutoring: Detecting Student Frustration and Disengagement Before They Happen \(/articles/affective-state/educational-tutoring\)](/articles/affective-state/educational-tutoring)
- [Governed Affective State for Compliant HR and Recruitment AI Agents \(/articles/affective-state/hr-recruitment-agents\)](/articles/affective-state/hr-recruitment-agents)
- [When an Agent Mood Cannot Be Explained Afterward \(/articles/affective-state/affective-state-stakes\)](/articles/affective-state/affective-state-stakes)

APPLICATIONS · SPECIFIC

- [Replika Alternative: Governed Affective State vs Reconstructed Emotion \(/articles/affective-state/replika\)](/articles/affective-state/replika)
- [Character.ai Alternative: Persistent Affective State vs Prompt-Defined Personality \(/articles/affective-state/character-ai\)](/articles/affective-state/character-ai)
- [Woebot vs Deterministic Affective State: Why a Therapy Chatbot Has No Persistent Modulation Field \(/articles/affective-state/woebot\)](/articles/affective-state/woebot)
- [Elomia vs. a Governed Affective State Field: What Persists Between Sessions? \(/articles/affective-state/elomia\)](/articles/affective-state/elomia)
- [Hume AI Alternative: Governed Affective State Beyond Emotion Measurement \(/articles/affective-state/hume-ai\)](/articles/affective-state/hume-ai)
- [Affectiva Alternative: Reading Human Emotion vs. Governing an Agent's Own Affective State \(/articles/affective-state/affectiva\)](/articles/affective-state/affectiva)
- [Cogito vs Governed Agent Affect: Reading Human Emotion Is Not an Agent's Internal State \(/articles/affective-state/cogito\)](/articles/affective-state/cogito)
- [Beyond Verbal vs Governed Affective State: Reading Emotion Versus Modulating Cognition \(/articles/affective-state/beyond-verbal\)](/articles/affective-state/beyond-verbal)

- [EmotiBit vs Governed Affective Modulation: Physiology Without a Policy-Bounded Affect Field \(/articles/affective-state/emotibit\)](/articles/affective-state/emotibit).
- [Realeyes Alternative: Governed Affective State vs Per-Session Emotion Measurement \(/articles/affective-state/realeyes\)](/articles/affective-state/realeyes).

HOW-TO GUIDES

- [How to Build a Therapeutic AI That Keeps Consistent Emotional Continuity Across Sessions \(/articles/guides/therapeutic-ai-emotional-continuity\)](/articles/guides/therapeutic-ai-emotional-continuity).
- [How to Give an AI Companion a Consistent Personality Across Sessions \(/articles/guides/consistent-personality-ai-companion\)](/articles/guides/consistent-personality-ai-companion).
- [How to Stop an AI Companion's Personality From Drifting Over Time \(/articles/guides/stop-ai-personality-drift\)](/articles/guides/stop-ai-personality-drift).

TERMINOLOGY

- [semantic hysteresis \(/glossary#semantic-hysteresis\)](/glossary#semantic-hysteresis)

[Affective State overview → \(/affective-state\)](/affective-state)