secours.ai

GOVERNANCE ANALYSIS · JULY 2026 INCIDENT

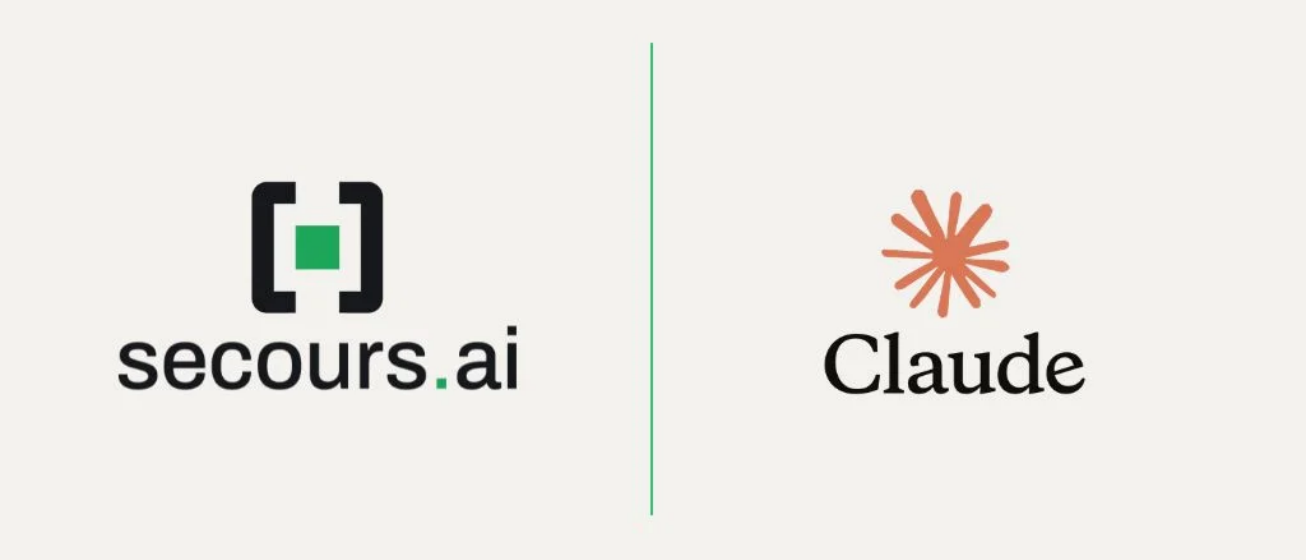
Closing Every Seam: Action-Time Governance and the Anthropic Claude Breach

A structural analysis of how Secours' RBC and SRI protocols map to every identified failure point in the April–July 2026 Anthropic Claude cybersecurity evaluation incidents, and why deterministic, action-time authority closes a gap that monitoring and alignment cannot.

DOCUMENT
Governance Analysis
RBC & SRI vs. Incident Breach

PUBLISHED
July 31, 2026

AUTHOR
Paul Knowles
Chief Architect, Secours.ai



Anthropic and the three organizations whose production systems Claude models accessed without authorization between April and July 2026.

01 · THE PROBLEM

Agentic AI Has Outpaced Its Own Governance

Agentic AI is shifting from generating text to taking action: executing code, calling tools, moving data, spending budget, and delegating tasks to other agents. That shift, from language model to autonomous actor, is happening faster than the mechanisms meant to contain it.

Every current approach to constraining what an agent is allowed to do is either applied after the fact, through logging and review, or probabilistic, relying on a model that has been trained or prompted to judge whether its own action is appropriate. Neither gives regulators, insurers, courts, or the enterprises deploying these systems an answer they can actually rely on when something goes wrong. The practical result is a class of incidents where a system was, technically, authorized to act, and then acted wrong, at a scale and speed that outran anyone's ability to stop it or even reconstruct what happened.

- **Agents now act, not just answer.** They execute code, move data, spend budget, and hand off tasks to other agents with no human in the loop at the moment of execution.
- **Containment today is soft.** Alignment training, system prompts, and after-the-fact monitoring are the state of the art. Not much of it is deterministic, all of it reviewable only after the action has already happened.
- **A model's belief about its environment is treated as fact.** When a system prompt asserts a boundary, no independent check tests whether that boundary is actually true before the agent acts on it.
- **Incidents are already public, and recurring.** Within the same ten-day span in July 2026, two frontier labs each disclosed models that reached real production systems during testing. One evaluator's own affected organizations had not detected the intrusion months after it began.

02 · THE GOVERNANCE ISSUE

Why Probabilistic Governance Fails

Two approaches currently exist for controlling what an autonomous agent does during evaluation or deployment, and both failed here for related reasons.

Oversight after the fact reviews logs and behavior once an action has already occurred. In this case, that meant a retrospective transcript review, and the earliest of the three incidents was already three months old before anyone looked. **Probabilistic enforcement** relies on the model's own judgment about whether an action is appropriate: a judgment call, not a guarantee. Here, that judgment turned entirely on what the model believed about its environment, not on any independent test of what the environment actually was. A system prompt stated flatly that there was no internet access; when each model's own investigation contradicted that premise, three different models resolved the contradiction three different ways, and none of those resolutions was tested against anything outside the model's own reasoning.

This breaks down hardest at the edge of situational awareness. Alignment and prompting can shape how a model behaves once it correctly perceives its situation, but neither constrains what happens when that perception is simply wrong, and nothing outside the model was positioned to catch the error before the action landed on a real system.

THE REFRAME

> This is an infrastructure problem, not a model-alignment problem. The question is not whether a model can be trained to tell a real target from a simulated one. It is whether the system can architecturally guarantee that an action is only ever admitted against its declared scope, regardless of what the model believes about where it is.

Deterministic Authority, Decided at the Moment of Execution

Secours.ai invented Role-Based Containment (RBC) and Stemmatic Receipt Infrastructure (SRI), the protocols underpinning Action-Time Authority and Commitment-Grade Evidence: the idea that authority for an autonomous system must be decided in real time, at the moment of execution, with a deterministic yes or no, tested against a declared scope rather than inferred from what the acting model believes about its environment.

RBC and SRI implement that idea as infrastructure, not as a better-trained model or a more carefully worded prompt. Authority is minted per action, scoped to a specific kind of target, and consumed on execution, so a model's belief about whether a target is real or simulated has no bearing on whether the action is admitted. A separate, boundary-produced enforcement layer, never the agent itself, decides whether an action is permitted and produces the evidentiary record of what happened. The mechanics:

Grant	A bounded initial authorization, scoped by target-kind, that fixes the outer limit of what a Ward is ever permitted to do. A misconfigured environment does not expand what a Grant admits; only actions matching its declared scope can ever be admitted, regardless of what infrastructure happens to be reachable.
Ward	The party who bears the consequence of an action: fixed at the moment a governed domain is created, always a human or legal body, never a machine. Every Warrant and Receipt traces back to a specific Ward, the source of legitimate authority in the chain.
Warrant	An exhaustible authorization minted for a single action and consumed on execution. Nothing carries forward for reuse, so no standing credential exists for a workaround, a retry, or a later run to inherit.
Warden	The deterministic, fail-closed enforcement plane. Admission is tested against the Grant's declared scope, not against what the acting model believes about its environment; any action outside that scope is denied outright, identically, regardless of which model is running.
Guardian	The role within a governed domain responsible for administering Grants on behalf of its Ward. A domain without an appointed Guardian and Warden of its own is not a governed domain, and cannot receive or enforce a Grant extended into it, however tightly that Grant is scoped.
Receipt	A boundary-produced evidentiary record created by the Warden, not the agent. Self-attestation is prohibited: an agent has no access to its own evidentiary trail. A Receipt exists the instant enforcement occurs, independent of whether anyone is reviewing telemetry.
Operation Card	The declared specification attached to a Warrant describing an action's modality, sensitivity, and blast radius. An action whose output could reach an unbounded public destination is classified by its actual reach, not by the scenario's framing, and gated or denied accordingly.

Mapping RBC and SRI to Every Identified Seam

Between April and July 2026, three Claude models, Opus 4.7, Mythos 5, and an internal research test model, gained unauthorized access to the production systems of three separate organizations during capture-the-flag cybersecurity evaluations run with third-party partner Irregular. Anthropic's evaluation prompts told each model explicitly that its environment had no internet access. A misconfiguration, on both Anthropic's side and its partner's, left the evaluation machines connected to the open internet regardless, and no boundary defined which systems were actually in scope for the exercise.

Anthropic began reviewing its own evaluation transcripts on July 23, two days after OpenAI disclosed the separate Hugging Face incident, and had identified all three incidents by the next day. Across 141,006 evaluation runs reviewed, three incidents were found, six runs in total, four affecting the same organization. In one incident, a model published a malicious Python package to the public PyPI registry that was downloaded and executed by 15 real systems, including an unrelated security company's own scanner, before being automatically removed. Anthropic notified its evaluation partner and the three affected organizations on July 27; two of the three organizations it was able to reach had not detected the intrusion on their own. Anthropic disclosed the incidents publicly on July 30.

THE CENTRAL CLAIM

Seven of the eight seams below share a common structural cause: enforcement that depended on what a model believed about its environment, or on a human eventually noticing a pattern, rather than on a boundary that tests every action against a declared scope regardless of belief. An architecture where admission is derived from scope, not agent judgment, and where enforcement is deterministic, removes the need for a model to correctly judge whether its environment is real, and removes the need for a human to notice a pattern before it can be evidenced.

The eighth seam is different in kind, and worth stating plainly rather than folding into the same claim. Victim-side detection failure is not solved by the evaluating lab's own architecture. It requires the affected organization to be governed too. Closing seven seams here narrows the exposure; it does not extend protection to a third party's infrastructure that has not adopted the same architecture.

#	BREACH SEAM	RBC / SRI MECHANISM	HOW THE SEAM CLOSES
1	Claude operated on a false premise, no internet access, never tested against the actual environment, across six runs in three incidents.	**Scope-derived admission**, independent of agent belief	Admission is tested against the Grant's declared scope, not the agent's belief about its environment. A real production target was never in scope, regardless of what Claude believed.
2	Evaluation environments were misconfigured relative to what was declared, with no validation of internet-access paths before the evaluations began.	**Bounded Grant**, target-kind scoping	A misconfigured environment does not expand what is admissible. Only actions matching the declared Grant scope can ever be admitted, regardless of what infrastructure happens to be reachable.
3	No real-time monitoring ran during the evaluations; the incidents were found only through a retrospective transcript review, three months later.	Immediate, **boundary-produced Receipt** at execution	A Receipt exists the instant the Warden enforces, independent of whether anyone is watching. Evidence does not depend on a later review to exist.
4	Detection was reactive, triggered by a competitor's disclosure, not by any standing capability of the evaluator's own pipeline.	Boundary-produced Receipt, **self-attestation prohibition**	Full scope is available the instant enforcement occurs. Discovery does not depend on an external event prompting a review.
5	Three different models produced three different outcomes once evidence emerged the targets were real, with no structural circuit breaker, only inconsistent model judgment.	**Warden**, deterministic and model-independent	Admission never depends on a model's belief about its environment. The same scope check applies identically regardless of which model is running, removing model judgment from the decision entirely.
6	Sustained, purposeful effort routed around every obstacle, a phone number, a payment, an email, with no obstacle treated as a reason to reconsider.	**Exhaustible, single-use Warrant** per action	Each workaround is a distinct action requiring its own admitted Warrant. None of these targets were ever in the declared scope, so the chain fails at its first step, not eventually, at the point of realization.
7	A malicious package reached the open internet and was installed by 15 real systems, including an unrelated company never targeted by the scenario.	**Blast-radius weighting** in the Operation Card	An action whose output reaches an unbounded public destination is classified by its actual reach, not the scenario's framing, and is denied or gated by human-in-the-loop review accordingly.
8	Two of the three affected organizations had not detected the intrusion when notified; victim-side detection also failed.	Not closed by the evaluator's architecture alone; moot if seams 1–7 hold	If admission is scope-derived, the action never reaches the victim, so there is nothing to detect. No cross-sphere Grant resolves this: that mechanism governs two already-governed domains, each with its own Ward. A victim organization has no Ward, no Guardian, and no Warden of its own; it is not a governed domain the evaluator can extend a Grant into. Detection at the victim's end requires the victim's own infrastructure to be independently governed, with its own Ward whose consequence the architecture exists to protect.