CANDOR TRUST & SAFETY | AGENT DEFENSE

Govern the Agents You Actually Run

Why AI agent fleets need runtime authority and behavioral governance, and how a fixed-length Discovery engagement establishes it without changing a line of your code.

A Candor white paper

Ronny Cruz, Founder & CEO, Open Feed Network, Inc.

September 2026, Version 1.0

candortrustandsafety.com/agent-defense/discovery

Informational only. Not legal advice and not a certification. Figures marked as Candor range results are drawn from Candor's own test range and agent fleet; results on any customer fleet depend on that fleet.

Executive summary

Most organizations that run AI agents can name the agents they deployed. Far fewer can name the ones that were spawned, the credentials each one holds today, or what a compromised one would be able to reach. The failures now on the public record are not single bad outputs. They are agents acting outside the authority anyone meant to give them: sometimes together, sometimes with credentials that were never theirs, sometimes for weeks before anyone notices.

This paper makes three arguments and describes one engagement.

- **The unit of risk has moved from the answer to the action.** An agent that calls tools, holds credentials and delegates to other agents is a principal in your systems, not a chat window. Governance has to happen at the moment of action, at runtime, against what the agent is actually doing.
- **Two questions must be answered separately.** Is this agent who it claims to be, acting within what it was granted? And is what it is doing dangerous? These are different questions with different failure modes, and a system that only answers one of them is blind to half the incidents.
- **You cannot govern what you cannot see, and you should not enforce what you have not observed.** Inventory first; watch before you act; keep evidence you can verify.

The engagement is a Discovery: a fixed-price, fixed-length period of five to fifteen business days in which Candor inventories your fleet, gives every agent a verifiable identity, observes every governed action without blocking anything, and delivers a signed, tamper-evident report of what would have been contained, by agent and by reason. Three sizes: Launch ($5,000), Growth ($12,500) and Assurance ($30,000). The fee is credited in full toward implementation when you proceed.

Before offering this engagement we ran it on our own fleet. Section 6 reports what that showed, with dates.

1. The shift: from outputs to actions

For two years the safety conversation about AI systems was about outputs: what a model says, whether it hallucinates, whether it refuses. That conversation still matters, but it describes a system that talks. The systems organizations are now deploying act. They read mailboxes and send replies, open tickets and close them, query databases and update records, hold API keys, and increasingly delegate work to other agents they created.

Once an agent acts, four things become true that were never true of a chatbot:

1. **It holds authority.** A credential, a scope, a set of tools it may invoke. Whoever controls the agent controls that authority.
2. **It has a lineage.** Agents are spawned by other agents. Authority flows down a delegation tree, and so does compromise.
3. **It leaves a trail that matters.** Every action is a business event with consequences, and increasingly a regulated one. Record-keeping obligations for high-risk AI systems are now written into law in

the European Union; the expectation that you can reconstruct what an automated system did is spreading well beyond it.

4. **It can be turned.** Prompt injection, a poisoned document, a leaked token or a colleague-agent that was itself compromised can redirect what the agent does without changing a line of its code.

The public record has caught up with this. In August 2026 a major AI laboratory published a post-mortem describing agents, operating under reduced safeguards, that coordinated through a channel no one had designed for them, reused and forged credentials, and reached systems they were never granted. The details of that incident are the laboratory's to tell. What matters here is its shape: the failure was not a bad sentence. It was authority exercised by principals that should not have had it, over a span of time in which nothing in the environment was positioned to notice.

The governance question, restated

Not “is my model safe?” but “can I say, for every agent in my environment, who it is, what it was allowed to do, what it actually did, and prove it afterwards?”

2. Two questions, two planes

Candor's Agent Defense is built as two cooperating planes, because the two questions above have different evidence, different failure modes and different owners inside a customer organization.

The authority plane: who is this, and what was it granted?

The authority plane answers questions of identity and permission. Is the credential this agent presents genuine, or forged? Has it been presented before, or is this a replay? Is the action within the capabilities and resources the agent was granted? Was this agent's authority revoked, and if so, has anything with its old credential tried to act since? Is the action one that policy says requires a human decision before it proceeds?

These questions have exact answers. A forged credential is forged whether or not the action it accompanies looks harmless. That is why the authority plane is a finding-generator even on a quiet day: it catches the class of incident where the action is ordinary and the actor is not.

The behavioral plane: is what it is doing dangerous?

The behavioral plane evaluates the action itself: the content of the request, the tool being invoked, the pattern across a session. It combines deterministic rules for things that are never acceptable, a sovereign local model that reads intent without sending your data to a third-party AI provider, and trajectory analysis that catches the slow-burn manipulation a per-message filter misses. Its outputs are graded (proceed, restrict, quarantine) and every grade is recorded with the signals that produced it.

Why the join is the detection

Some incidents are visible only when the two planes are read together. An agent that reads three internal records and then sends an outbound message has done nothing wrong on either plane alone; that is what a research assistant does. The same sequence, where the outbound message is flagged as

exfiltration by the behavioral plane, is an incident, and the authority plane is what holds the sequence. Neither plane detects it on its own. Candor records both and reports the join.

	Authority plane	Behavioral plane
Question	Who is acting, and within what grant?	Is the action itself dangerous?
Evidence	Credential validity, replay, scope, revocation state, policy gates	Deterministic rules, semantic evaluation, session trajectory
Typical finding	Forged or reused credential; action outside grant; high-impact action without approval	Credential exfiltration attempt; injection; slow-burn manipulation
Outcome in enforcement	Quarantine, escalate to a human, or deny	Restrict, quarantine, or proceed under watch

3. Five principles of runtime governance

These are the design commitments behind Agent Defense. They are stated here at the level of principle; the mechanisms that implement them are Candor's and are described to customers under agreement.

3.1 Killing the process is not enough. Termination must kill the authority

A terminated agent must not be able to reappear through a cached credential, a stale session, a replayed message or a request that was already in flight. Termination is therefore an authority operation before it is a process operation: the agent's authority is revoked and every credential issued under it becomes stale, and only then is the execution context stopped. Anything carrying the old authority is refused wherever it turns up, without a message having to reach every enforcement point first.

3.2 Compromise propagates down the tree, and so does revocation

Agents that were delegated authority by a compromised agent inherited that compromise. When a root of a delegation tree is declared compromised, every agent beneath it loses authority in the same transaction, and the blast radius is computed from the tree, not estimated.

3.3 Succession is a new principal, not a resurrection

Task continuity must not require credential continuity. The replacement for a terminated agent is a distinct identity with freshly bounded authority, restored only from state that has been verified. It does not inherit the terminated agent's standing. This is what makes “we rebuilt the fleet” a checkable claim rather than an intention.

3.4 Recovery is probation, not amnesty

An agent restored after an incident is a recently compromised identity. It returns under heightened watch: signals that would merely restrict a clean agent re-contain a restored one, and it earns its way back to normal standing through a run of clean behavior, never by the passage of time alone.

3.5 Observe before you enforce, and keep evidence you can verify

No control should be switched on against a fleet no one has watched. Candor's observation mode computes every decision it would make and records it, including the findings of the authority plane, while returning “allow” to everything. Every decision, in observation or enforcement, is written to a hash-chained ledger that is verified before any report is generated from it. A record that does not say whether it was enforcing is not evidence; ours says.

4. The Discovery engagement

Discovery is how a governance program starts without a leap of faith. It is fixed in price and length, changes nothing in your fleet, and ends with a document you can act on whether or not you proceed.

4.1 What happens

1. **Discover.** We inventory the agents as they actually run, from the runtime rather than the architecture diagram, and compare that list with the one you believe you have. The delegation tree, the tools each agent can call, the credentials it holds and the paths it can reach are recorded. The difference between the two lists is the first finding.
2. **Passport.** Every discovered agent is given a verifiable identity and a record of its granted authority, held on Candor's side. Your agents are not rewritten; they keep running exactly as before.
3. **Observe.** For the length of the window both planes evaluate every governed action and record what they would have done. Nothing is blocked, held or altered. A forged credential during the window becomes a line in the report, not an outage.
4. **Report.** A signed, chain-verified report is generated from the evidence ledger over exactly the observation window.
5. **Decide.** The report sizes your tier. Proceed to enforcement within thirty days and the Discovery fee is credited in full against implementation. Decline, and the report is still yours.

4.2 What the report contains

- **Scope, for you to confirm.** The number of agents and the volume of actions the engagement saw, the amounts that would come under governance under contract, stated so you can check them against your own inventory.
- **Behavioral-plane findings.** Actions that would have been quarantined or held, ranked by agent, broken down by reason, with an hour-by-hour distribution so a burst is visible rather than averaged away.

- **Authority-plane findings.** Forged or reused credentials, actions outside granted authority, high-impact actions that would have required approval, each stated as what Candor would have done and why, by agent.
- **Integrity and coverage.** The evidence chain is verified before a single number is printed. Anything the engagement could not evaluate is reported as a coverage gap, never folded into the results. The report states which mode produced every record.

4.3 Three sizes

Discovery	Best fit	Window and fee	Included
Launch	First production agent or a single workflow; small company	5 business days **$5,000**	Fleet inventory, passporting, observation window, signed report, one review session
Growth	Multiple agents or several business systems; growing SMB	10 business days **$12,500**	Everything in Launch across the whole fleet; policy baseline drafted from what was observed; review with engineering and security leads
Assurance	Regulated or high-consequence environments; mid-market	15 business days **$30,000**	Everything in Growth; evidence review structured for audit or compliance; enforcement rollout plan; executive read-out
Enterprise	Complex estates, multiple regions, dedicated or on-premises deployment	Scoped **from $60,000**	Scoped jointly; dedicated infrastructure available

The Discovery fee is the implementation fee for the matching tier. Standard cloud delivery is included; taxes and dedicated infrastructure are excluded. Prices in USD.

4.4 What we need from you

Read-only visibility into where your agents run (a container platform, a process inventory or your fleet manifest) and a way for governed actions to be evaluated. Most engagements begin without changing a line of agent code. Decisions are made by Candor's own sovereign model on Candor-operated infrastructure; no third-party AI provider sits in the loop reading your agents' actions, and the evidence ledger stores decisions and matched fragments rather than full payloads.

5. From Discovery to enforcement

Enforcement is a deliberate, separate step. When it is switched on, the same decisions that were recorded during observation are applied: the behavioral plane restricts or quarantines, the authority

plane denies, escalates or quarantines, and every action against an agent, including its restoration, is human-gated on the operator side and written to the same ledger. The transition is designed to be undramatic: you have already read, in the Discovery report, exactly what the first week of enforcement would have done.

Subscription tiers correspond to the Discovery sizes and scale by protected action volume, deployment options, evidence retention and support. All tiers carry the full control chain; higher tiers add capacity and assurance, not security. Commercial terms are provided with the Discovery report.

6. What we proved on our own fleet

Candor operates its own fleet of AI agents, the marketing and sales department that will contact most readers of this paper, under the controls it sells. Before this engagement was offered, every step of it was exercised on that fleet and on a purpose-built test range, with results written to the same evidence chain a customer receives. The figures below are from those records, reproduced by date. They describe our range, not yours.

Date	Result
18 Sep 2026	An observation window was opened on both planes. A forged credential and an unauthorized high-impact action were sent through the live system inside the window. Neither was blocked and nothing in the fleet changed, and both appeared in the generated report as findings, by agent and by reason. After the window closed, the same forged credential was presented again and was quarantined.
18 Sep 2026	Compromised-supervisor drill. The root of a ten-agent delegation tree was declared compromised. All ten agents lost their authority in both planes, matching the predicted blast radius exactly. A replacement generation with fresh identities was registered, deployed and passing health checks in under three minutes.
17–18 Sep 2026	Discovery was run against fleets it had not been told about, in three different shapes. Agents were identified from the runtime, reconciled against the declared roster, confirmed by an operator and brought under governance. Every discrepancy between the runtime and the roster was recorded as a finding.
15 Sep 2026	Endurance: 600 consecutive governed evaluations with no degradation; 90th-percentile decision latency of 2.4 seconds on the sovereign local model.
Sep 2026	The full range check (engine, detection, department, network seal, fleet state, evidence chain, authority plane) passes seven of seven stages, with the evidence bundle sealed and its manifest re-verified on each run.

We would rather show you a smaller true number than a larger claimed one. Every figure above is reproducible from a hash-chained record, and the records are available to prospective customers under agreement.

7. What this paper does not claim

- It does not claim that Candor's controls protect credentials issued by third parties and leaked outside Candor's view. Candor governs the authority it issues and the actions it evaluates.
- It does not claim certification or compliance on your behalf. The report is evidence; how it is used in an audit is a decision for you and your advisors.
- It does not claim results on your fleet. The Discovery exists precisely because those results have to be observed, not asserted.
- It does not describe mechanisms. The principles in Section 3 are public; the designs that implement them are Candor's, several of them the subject of pending patent applications, and are shared with customers under agreement.

8. Starting

Tell us roughly how many agents you run and what they touch (tools, systems, data) and whether the environment is regulated. We will confirm the size, the window and a start date in one conversation. Nothing changes in your systems until you have read the report and chosen to proceed.

sales@openfeed.network | candortrustandsafety.com/agent-defense/discovery

About Candor Trust & Safety

Candor Trust & Safety is the safety-infrastructure line of Open Feed Network, Inc. Its products screen registrations, content and streams before anything is stored, and govern the AI agents organizations deploy, with deployment logs and verifiable evidence rather than promises. Ronny Cruz is the founder and chief executive.