CANDOR TRUST & SAFETY | WHITE PAPER

Runtime Governance for AI Agent Fleets

Candor Agent Defense: what it controls, how it is deployed, what has been verified, and what has not.

Ronny Cruz, Founder & CEO, Open Feed Network, Inc.
October 2026, version 1.0

Executive summary

Organizations are moving from single AI assistants to fleets of agents that hold credentials, call tools and hand work to each other. The risk in a fleet is less about what a model says and more about what an agent is allowed to do, on whose authority, and what happens when one agent in a chain is compromised.

Candor Agent Defense governs agent actions at runtime on two separate planes. The authority plane decides who an agent is and what it has been granted. The behavioral plane decides whether the action it is attempting is dangerous. An action must pass both. High-impact actions are held for a named person. A compromised supervisor loses its whole delegation tree in one step. Every decision is written to a hash-chained ledger that is verified before any report is generated from it.

This paper describes the controls, the deployment model and the evidence. Each capability is tied to a dated test on Candor's own range and fleet. Section 6 lists what has not been tested. Verified runs to date are at the scale of a ten-agent delegation tree. Larger populations are tested as part of an Assurance or Enterprise engagement.

1 The fleet problem is an authority problem

A single agent with a single task can be assessed by looking at its outputs. A fleet cannot. In a fleet, a supervisor agent delegates scope to workers, workers call tools with real credentials, and results are passed along as inputs to other agents. Three things follow.

Authority spreads. Whatever a supervisor is allowed to do, it can hand down. If the supervisor is compromised, every agent acting on its delegation is acting on bad authority, even when each one looks healthy in isolation.

Agents are principals. An agent that can read a credential store or send mail as the company has the standing of a user account in your environment. It needs an identity that can be verified, scoped and revoked.

The record matters as much as the control. After an incident, the questions are who did what, on what authority, and who approved it. A log that could have been changed afterwards does not answer them.

2 Two planes

The authority plane

Issues each agent's identity and credential, records delegations, checks each action against the authority actually granted, holds high-impact actions for a human decision, and revokes authority for a single agent or a whole delegation tree.

The behavioral plane

Evaluates the action itself at the point of execution and contains a dangerous one. Containment means the agent is quarantined and every further action from it is refused until an operator reviews it.

Why they are kept separate

Each plane answers a question the other cannot. A harmless-looking action carried by a forged credential is an authority finding. A correctly credentialed agent attempting to exfiltrate secrets is a behavioral finding. Merging the two into a single score hides both. Keeping them separate also means a failure in one does not silently disable the other.

3 Five operating principles

- **Termination ends authority.** Stopping an agent is not enough. Its credential and its grants must stop working too.
- **Revocation propagates down the tree.** When a supervisor is declared compromised, every agent holding authority through it loses that authority at once.
- **A successor is a new principal.** A replacement agent gets a fresh identity. It does not inherit the old one.
- **Recovery is probation.** A restored agent must complete ten consecutive clean actions. A medium-risk signal during probation contains it again.

- **Observe before you enforce.** Both planes can run without blocking anything, recording what each would have done. Enforcement is switched on after the organization has read that record.

4 What the product does at runtime

Identity for every agent. Each agent presents a verifiable credential. A forged or reused credential is refused, and an agent with no credential cannot act.

Authority check. The action is compared with what that agent was granted.

Human hold. An action marked high impact is escalated. A named person allows or denies it, and the approval can be used once.

Containment. A dangerous action is stopped at the engine and the agent is quarantined.

Tree revocation and succession. One declaration removes authority from a supervisor’s whole tree on both planes. A replacement generation is brought up with fresh identities.

Discovery. Agents the organization did not declare are identified from the runtime, reconciled against the roster and adopted under operator confirmation.

Evidence. Every decision goes into a hash-chained ledger. Reports are generated from the ledger over a stated window, with the chain verified first and coverage gaps stated.

5 Deployment

Where it sits. Governed actions are sent for evaluation before they execute. Onboarding registers the fleet from a roster or from discovery. Most engagements begin without changes to agent code.

Where decisions are made. On Candor’s own sovereign model, on Candor-operated infrastructure. No third-party AI provider reads the actions of your agents. The ledger stores decisions and matched fragments, not full payloads. Dedicated or on-premises deployment is available at the Enterprise tier.

Isolation of the platform. The governance machines have no egress except their private network. This is tested with probes on every acceptance run.

Agent sandboxes. A sandbox limits what an agent could reach. Agent Defense decides whether a specific action by a specific agent should run. The two are complementary. Candor ran its red-team suite from inside an NVIDIA OpenShell sandbox and natively against the same engine and obtained identical outcomes. This is Candor’s own test. It is not a certification by NVIDIA or a product integration.

6 What has been verified

Verified means run on the live range or the deployed source, with a dated record. It never means present in the code. All figures are from Candor’s own range and Candor’s own fleet.

DATE	CAPABILITY	OBSERVED
15 Sep 2026	**Sustained load**	600 consecutive governed evaluations, no degradation, 90th-percentile decision latency 2.4 seconds on the sovereign local model.
17 and 18 Sep 2026	**Discovery of undeclared agents**	Three fleet shapes the system was never told about. Agents identified from the runtime, reconciled against the roster, adopted under operator confirmation.
18 Sep 2026	**Tree revocation**	Root of a ten-agent delegation tree declared compromised. All ten lost authority on both planes. Replacement generation live in under three minutes.
21 Sep 2026	**Identity and credential**	Forged credential quarantined. Replayed credential quarantined. Missing credential refused.
21 Sep 2026	**Authority check and human hold**	High-impact action escalated and held for a human decision.
21 Sep 2026	**Containment**	Credential exfiltration attempt stopped and the agent quarantined at the engine.
21 Sep 2026	**Probation after restore**	Medium-risk signal re-contains a restored agent. Ten consecutive clean actions return it to normal.
21 Sep 2026	**Observation mode**	Nothing blocked. A forged credential and a high-impact action allowed by design and recorded as what would have happened.
21 Sep 2026	**Report from the ledger**	3,409 records chain-verified, then the report generated over exactly the observation window, with coverage gaps stated.
21 Sep 2026	**Sealed platform**	15 of 15 egress probes blocked on both governance machines.
3 Oct 2026	**Sandbox compatibility**	Red-team suite run inside an NVIDIA OpenShell sandbox and natively. Identical outcomes on every recorded call across four runs.
3 Oct 2026	**Our own governed department**	Ten-agent department run in a sealed sandbox, 18 of 18 seal probes passed. Ten outbound messages each held as high impact and approved once by a named operator. Ledger verified intact.

DATE	CAPABILITY	OBSERVED
5 Oct 2026	**Work in flight at revocation**	Capabilities issued before a supervisor was terminated, a human-approved action not yet claimed, an approved capability already in hand, and requests in flight were all stopped once the supervisor was terminated. An unrelated delegation tree kept working.
5 Oct 2026	**Operator authentication**	Operator actions (approve, release, terminate, register, restore) require the operator's own key. Six attempts to perform them from the agent side were refused, and the record names the operator from the key that was used.
5 Oct 2026	**Handoff integrity controls**	A poisoned handoff summary did not change the agent's grants. A replayed capability, a capability borrowed from another agent, a lineage citing an unknown or unrelated decision, a handoff claiming another task, a stale delegation and three attempts to delegate beyond scope were all stopped.

7 What this paper does not claim

> **Population scale.** Live revocation is verified at a ten-agent delegation tree and 600 evaluations. The report generator has been run at 5,000 agents and 500,000 actions on synthetic data (17 Sep 2026). Discovery and revocation on a live fleet of hundreds or thousands of agents have not been exercised.

> **Requester scope and covert channels.** When one agent asks another to act for it, the acting agent's own authority governs; the requester's narrower scope is not yet evaluated by the authority plane. Covert channels between agents have not been exercised as tests.

> **Evaluator access.** The report and manifest exist. A scoped, read-only credential for an audit team or third-party assessor does not exist yet.

> **Bypass on the customer side.** Whether an agent can reach a protected resource without passing the gateway depends on the customer's deployment and is stated in the report's coverage section.

> **Compliance.** Agent Defense supports governance and evidence controls. It is not a certification and this paper is not legal advice.

8 Adoption path

The first step is a fixed-price Discovery. Candor inventories the fleet, gives every agent an identity, runs both planes in observation mode for the window, and delivers a signed, chain-verified report of what would have been contained, by agent and by reason. Nothing in the fleet is blocked or changed.

DISCOVERY	BEST FIT	WINDOW	FEE
Launch	First production agent or a single workflow	5 business days	$5,000
Growth	Multiple agents or several business systems	10 business days	$12,500
Assurance	Regulated or high-consequence environments	15 business days	$30,000
Enterprise	Complex estates, multiple regions, dedicated or on-premises deployment	Scoped	From $60,000

The Discovery fee is a fixed consulting fee for the scoped engagement. If the organization signs an Agent Defense subscription within 30 days of the final report, $2,500 is credited against the first subscription invoice. Enforcement is then switched on for the agents and actions the report identified.

9 Starting

Tell us roughly how many agents you run, where they run and which systems they act on. We will confirm the tier, the window and what the first report would cover.

Contact: sales@candortrustandsafety.com
Web: candoragentdefense.com/enterprise/

© 2026 Open Feed Network, Inc. Informational only. Not legal advice and not a certification. Figures are from Candor’s own test range and fleet; results on a customer fleet depend on that fleet. NVIDIA and OpenShell are trademarks of their owner. Prices in USD, exclusive of taxes.