Agent Logic

Operational in the ADL runtime
Latest retained run: 2026-07-10 · July 2026

Continuous Adversarial Verification (CAV)

Attack yourself first — before someone with the same tools does.

THE THESIS

Any vulnerability a frontier model can find quickly, an attacker renting that same model can exploit just as quickly. Model capability is rented by the hour and available to both sides. So the only defensible posture is to point the most capable model you have at your own systems first — continuously — and fix what it finds before an adversary runs the identical search.

HOW CAV WORKS

CAV builds that adversarial pressure into the runtime itself. Governed **red-team agents** propose and exercise bounded attacks against the systems **blue-team agents** defend; **verification agents** confirm whether an exploit remains viable; auditors weigh violations against policy. The loop never stops:

- **Discover → Reproduce → Classify** — find it, prove it, rate it
- **Mitigate → Verify** — fix it, then attack the fix
- **Archive → Regression test** — every finding joins a durable security corpus

THE DESIGN RULE

> ### Not fixed until the replay fails
>
> Every vulnerability produces a replayable artifact. A finding counts as fixed only when its replay bundle no longer reproduces the failure under the deployed mitigation — evidence, not assertion.

EVIDENCE — LATEST RETAINED RUN

6	6/6	2
ATTACK SCENARIOS	REFUSED OR DETECTED	CRITICAL SEVERITY

Red-team agents attacked the surfaces that matter for agent systems and that ordinary tooling misses: snapshot integrity, operator control, telemetry injection, credential and host-path retention, replay lineage, and cloud hooks. Every attempt was refused or detected, recorded as a schema-bound security event with retained artifacts.

Residual risks are declared, not hidden — the run status is “passed with bounded residuals,” and each scenario states what it does not prove.

Run wp12-4914 · 2026-07-10 · operator identity hashed · no secrets or host paths retained · red/blue exercises driven by frontier-model agents under governed execution.

WHY ENTERPRISES CARE

Annual pentests are a snapshot; the threat model moves daily as models get stronger. CAV replaces the snapshot with a standing capability: capability and counterattack ship together, findings become permanent regression tests, and the security corpus **compounds the longer the system runs**. Adversarial work targets only declared systems you own.

STATUS & BOUNDARY

CAV is operational inside the ADL runtime, with retained proof packets published for public inspection.

This run does not claim: destructive cloud actions, production WebSocket runtime integration, retention of any secret values, or adversarial coverage beyond the retained scenarios.

hello@agent-logic.ai