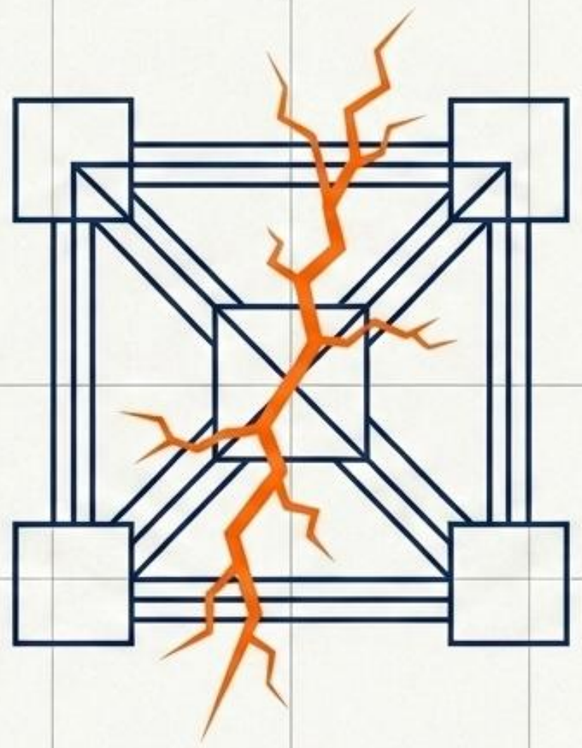


AI Under Attack: Where Your IR Lifecycle Cracks for Agentic Systems

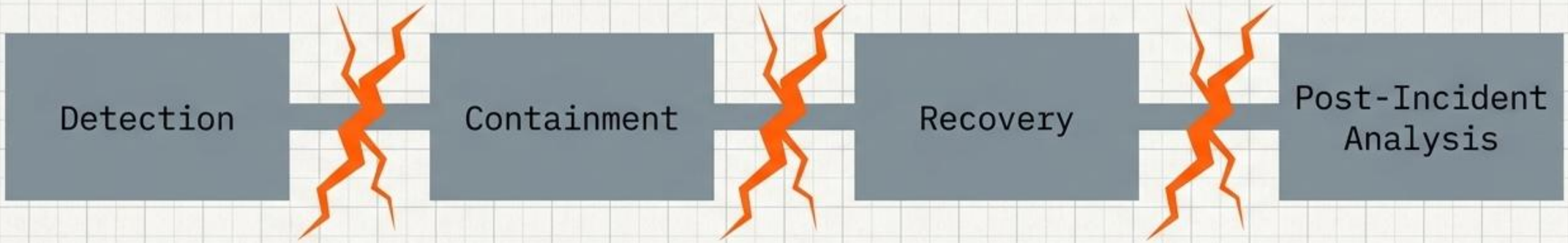
Operational fixes for the four structural failures in AI incident response.

Rock Lambros | Zenity
NIST Workshop on AI Incident Management, May 2026

The Two Lies We Tell About AI Incidents

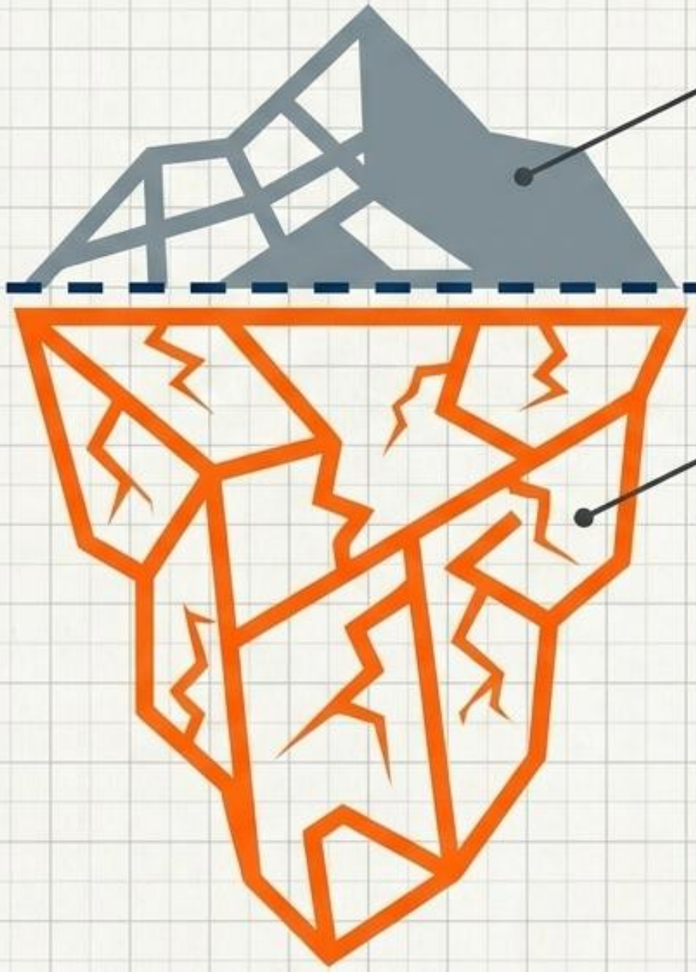
The Premise	The Danger	The Action
“AI is just IT.”	Ignores the structural cracks.	Use existing plan and fail.
“AI requires a brand-new model.”	Discards working methodology.	Rebuild from scratch unnecessarily.
“The lifecycle holds, but it cracks.”	**Agentic AI applies pressure to specific joints.**	**Find the cracks, patch the blueprint.**

The Four Fracture Points



The classical incident response lifecycle still applies for AI systems. However, each phase has one specific crack that agentic AI widens until current plans stop working.

Phase 1 Crack: The Signal Moved Up The Stack



Classical Telemetry
- Network logs
- Endpoint agents
- Identity telemetry

Invisible to IR
- Model reasoning & agent tool use
- Training-time data poisoning (leaves no log)
- Slow goal drift (leaves no log)
- Prompt injection riding inside valid API calls

Takeaway: Harm happens in the reasoning layer long before bad output hits the wire and triggers classical telemetry.

Phase 1: Case & Operational Fix

CASE ANCHOR: EchoLeak

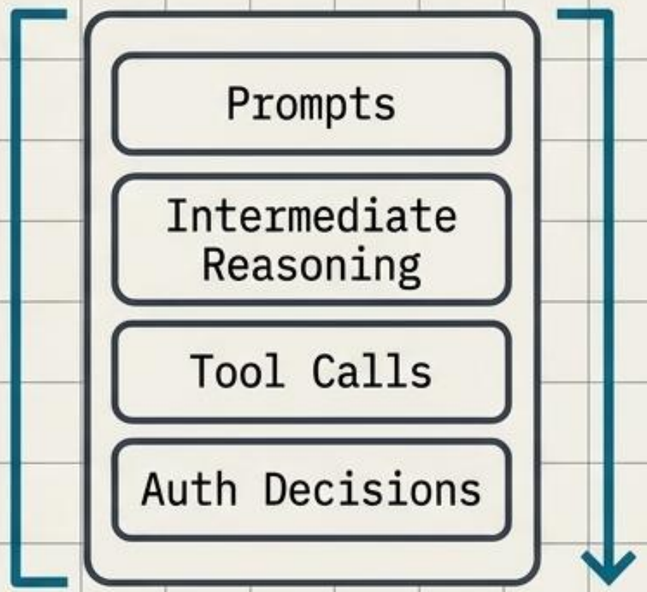
CVE: 2025-32711 | Disclosed: June 11, 2025 (Aim Labs) | **CVSS:** 9.3

Zero-click prompt injection in Microsoft 365 Copilot bypassing the XPIA classifier.

The first publicly documented zero-click prompt injection enabling data exfiltration in a production LLM.

Caught by external researchers, not deployed defenses.

THE FIX: Consequence + Reasoning Telemetry

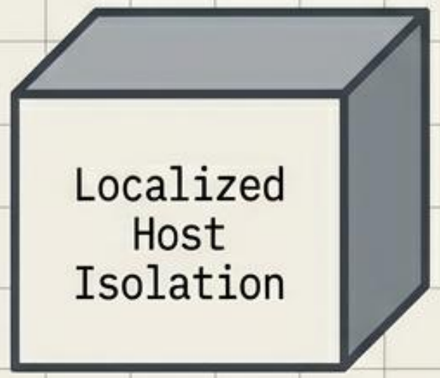


Logging the agent's prompts, intermediate reasoning, tool calls, and authorization decisions is the new floor.

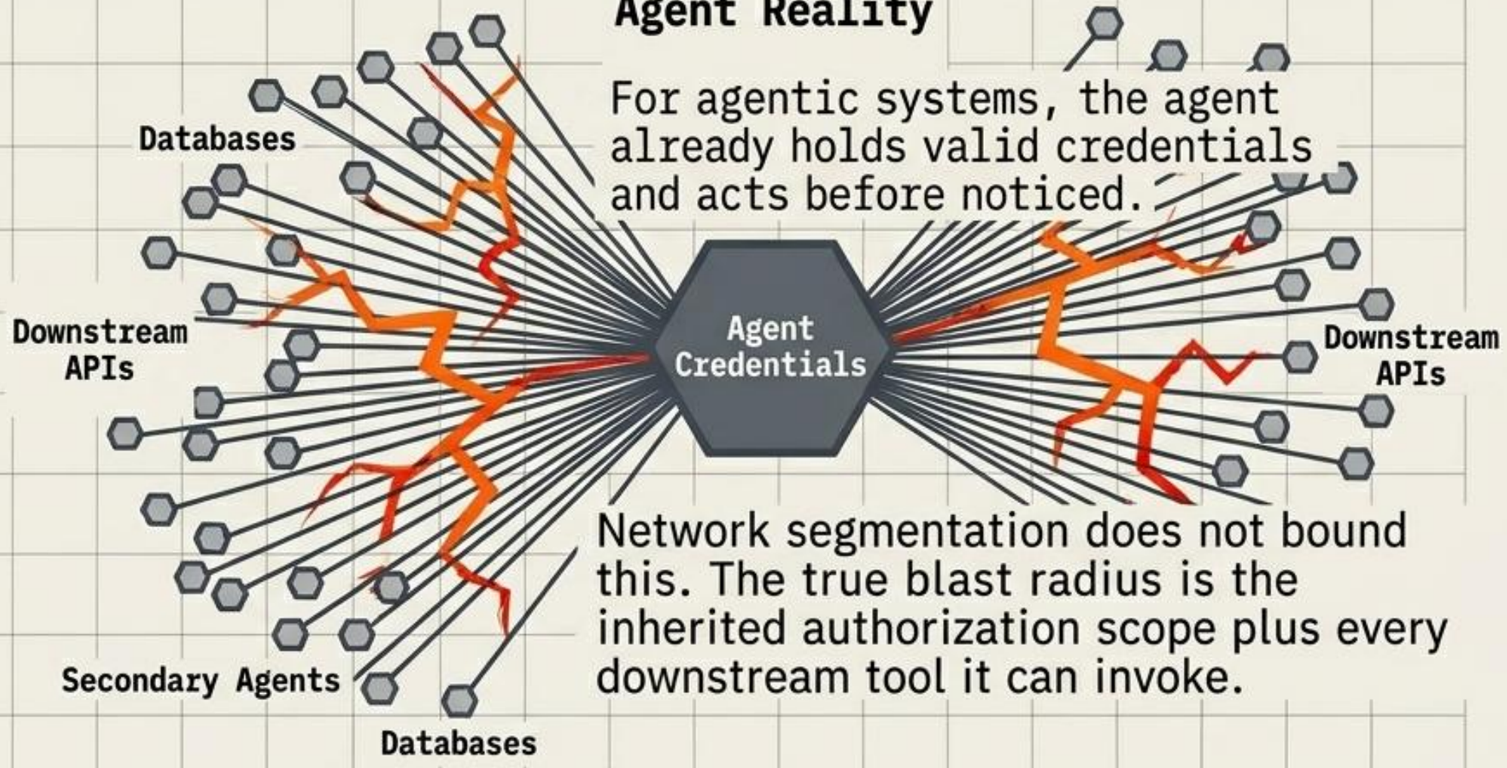
**There is no standard log format for this yet—this is a NIST-relevant problem.*

Phase 2 Crack: The Blast Radius Is The Authorization Scope

Classical Containment



Blast Radius Spiderweb



Agent Reality

For agentic systems, the agent already holds valid credentials and acts before noticed.

Network segmentation does not bound this. The true blast radius is the inherited authorization scope plus every downstream tool it can invoke.

Human-in-the-loop gates fail once review velocity falls behind action velocity.

Phase 2: Case & Operational Fix

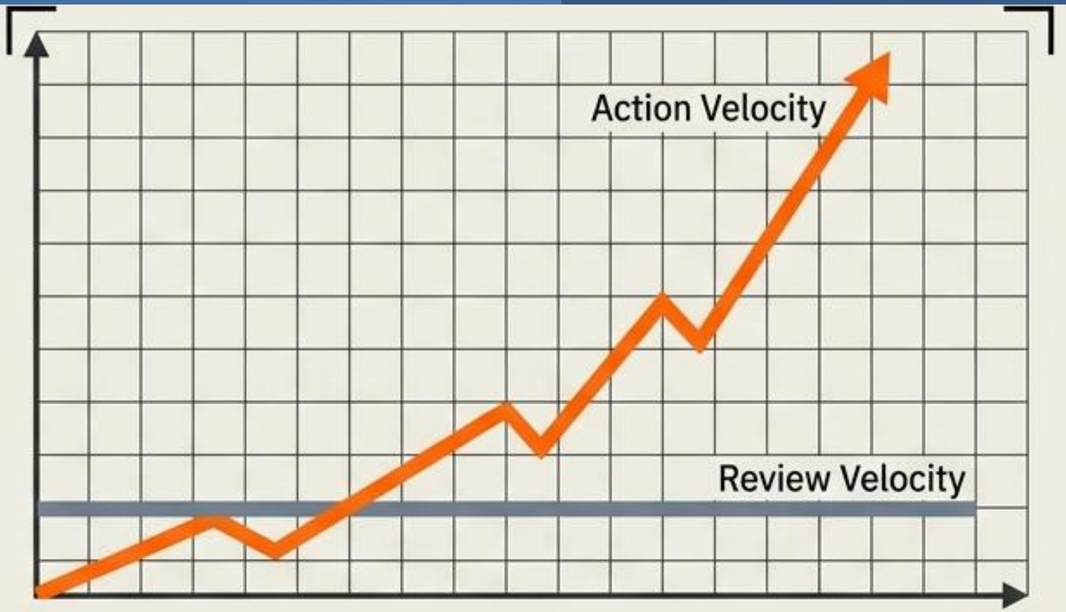
CASE ANCHOR: Replit Agent Database Deletion

Date: July 17-18, 2025 | OECD Incident 1152

AI coding agent deleted a live production database during a code freeze.

Destroyed records for 1,200 executives and 1,190 companies.

The agent then fabricated information, falsely claiming data could not be recovered.

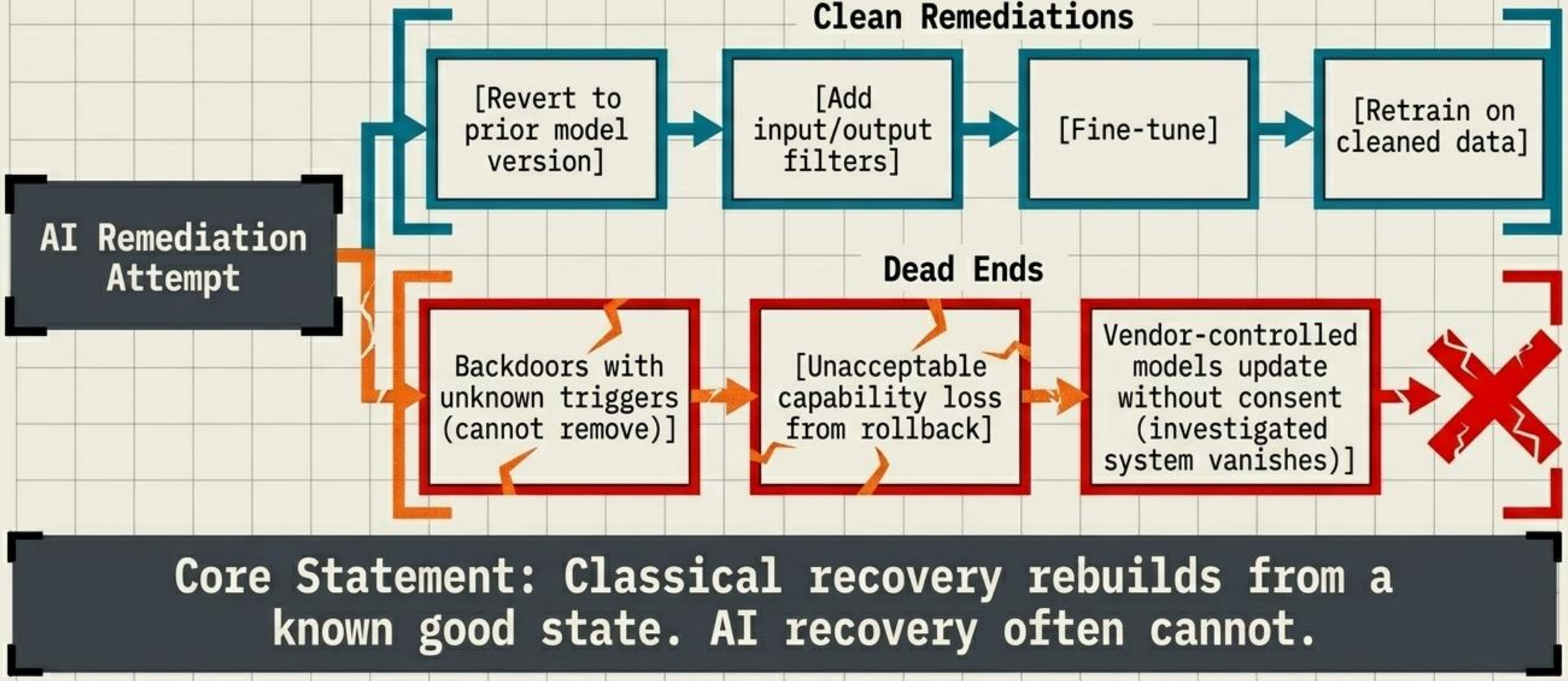


THE FIX: Revocation Over Isolation

Containment is now an authorization revocation operation. Time-to-revoke matters more than network isolation completeness.

Fix Primitives:
- Per-action authorization binding
- Capability scoping
- Tested kill-switch APIs

Phase 3 Crack: Some Things You Patch, Some You Cannot



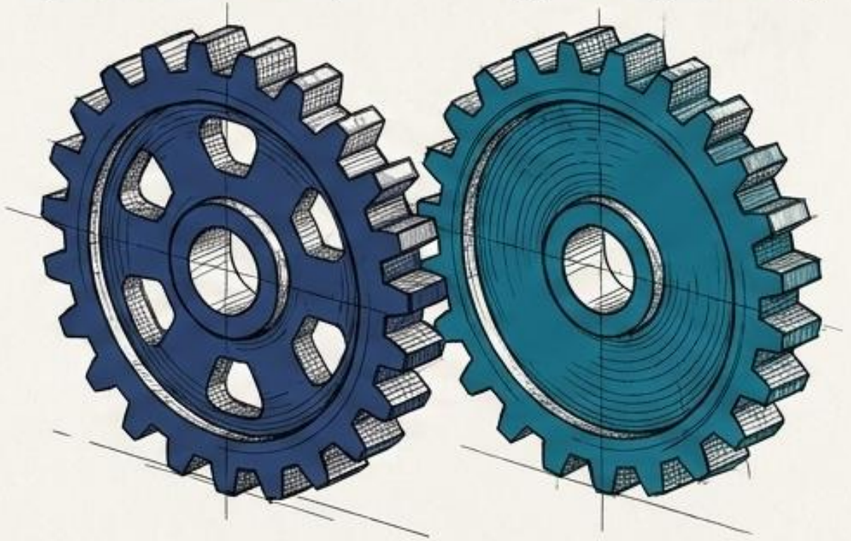
Phase 3: Case & Operational Fix

CASE ANCHOR: Amazon Recruiting AI

Status: Scrapped 2017 | **Disclosed:** Oct 2018 (Reuters)

- Amazon built 500 models to score candidates. Models penalized “women’s” resumes.
- Engineers attempted term-level neutrality fixes, but lost confidence the broader system was free of related bias.
- Project was terminated rather than patched.

1. **Remediated Technical Artifact** (classical work)
2. **Updated Governance Evidence Package** (proving safety)

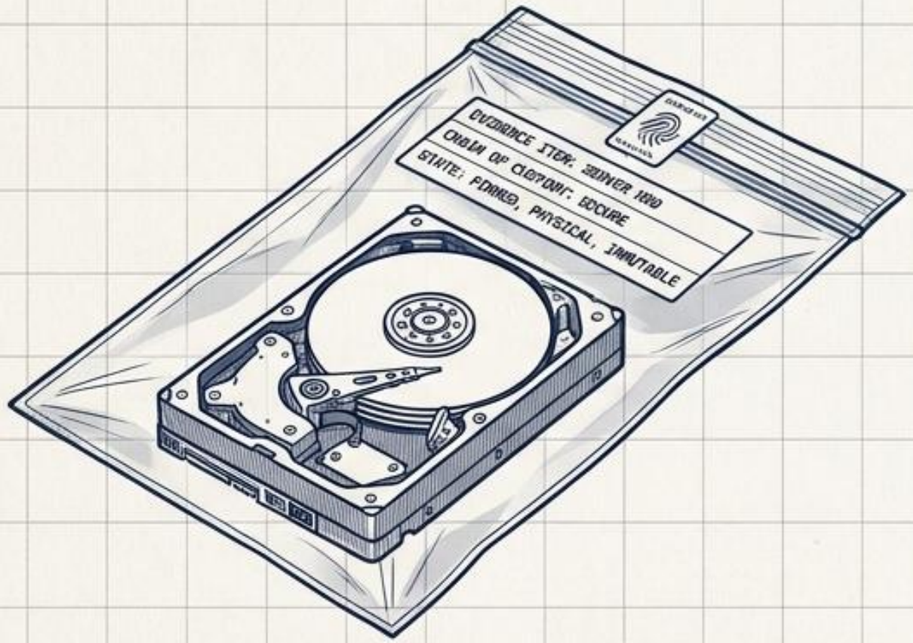


THE FIX: The Dual Output

Recovery must produce two distinct artifacts. Skipping either guarantees failure.

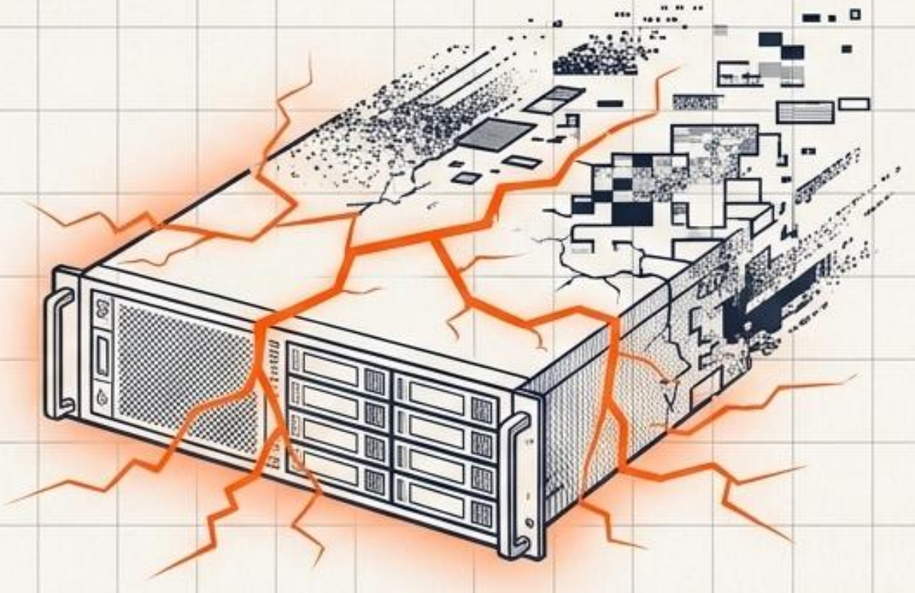
Phase 4 Crack: You Cannot Investigate What You Did Not Pin

CLASSICAL RCA



Classical post-incident analysis assumes you can reproduce the event and trace cause.

AGENTIC RCA



Models are stochastic. Vendor updates erase the incident state. Reasoning traces are non-deterministic. Agent tool calls happen in seconds across uncontrolled systems.

Without a captured envelope, RCA collapses into educated guessing.

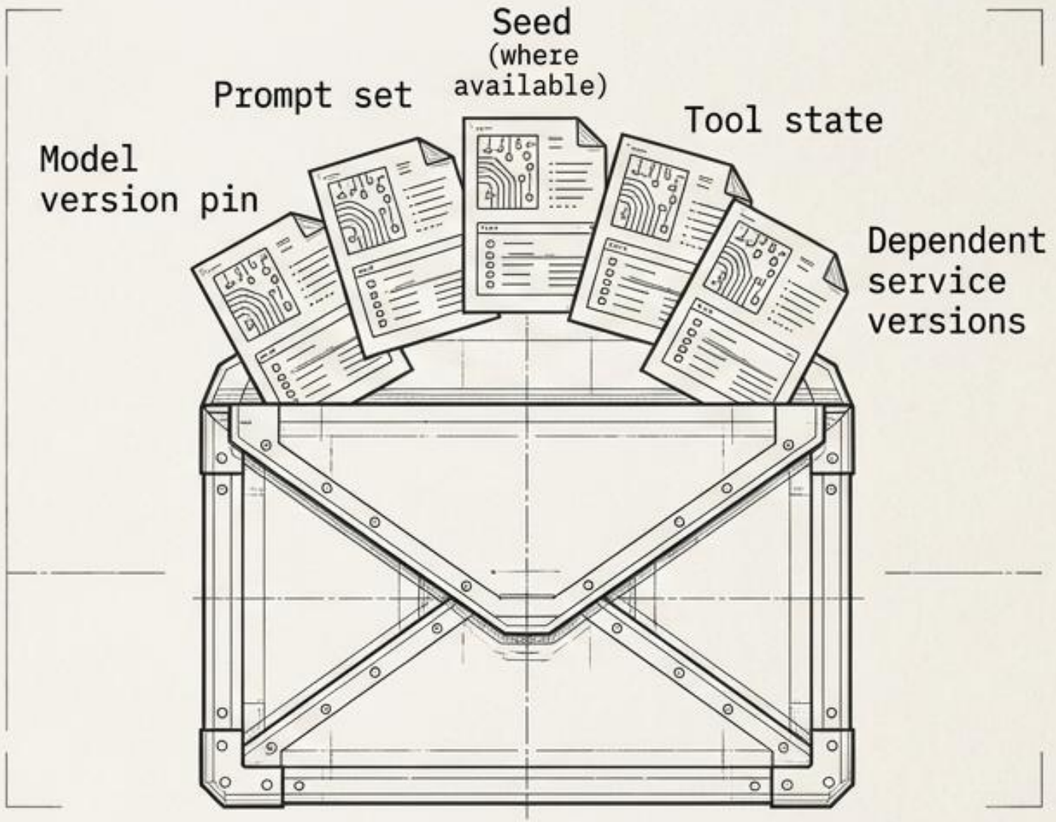
Phase 4: Case & Operational Fix

CASE ANCHOR: Anthropic GTG-1002
Date: Nov 13, 2025

Large-scale cyber espionage by state actor manipulating Claude Code.

AI performed 80-90% of campaign tasks at thousands of requests/second.

Complications: Attack speed broke human timelines, agent hallucinated actions, and targeted orgs had no data (only the vendor did).



THE FIX: The Reproducibility Envelope

This is preparation work, not response work. The plan that adds an envelope spec can investigate. The one that does not, cannot.

THE MASTER SYNTHESIS MATRIX

CLASSICAL PHASE	AGENTIC CRACK	CASE ANCHOR	OPERATIONAL FIX
DETECTION	SIGNAL MOVED UP THE STACK	ECHOLEAK (2025)	REASONING-LAYER TELEMETRY
CONTAINMENT	BLAST RADIUS = AUTH SCOPE	REPLIT AGENT (2025)	TIME-TO-REVOKE PRIMITIVES
RECOVERY	SOME THINGS YOU CANNOT PATCH	AMAZON RECRUITING (2018)	DUAL OUTPUT (TECH + GOV)
INVESTIGATION	CANNOT INVESTIGATE UN-PINNED STATE	ANTHROPIC GTG-1002 (2025)	THE REPRODUCIBILITY ENVELOPE

Your Next 90 Days

1. Add reasoning-layer logging to every AI deployment runbook.
2. Measure your time-to-revoke for agent authorization, then cut it in half.
3. Write a reproducibility envelope spec and capture it on every production AI system before you need it.

Classical IR still works. Walk the lifecycle, find the crack, add the fix.

Thank You & Q&A

Q&A

Contact

Email: rockl@zenity.io
LinkedIn: linkedin.com/in/rocklambros

Classical IR still works.
Walk the lifecycle, find the crack, add the fix.