ISSUE 003 · AUGUST 2026

THE AI RUNTIME REPORT

Runtime Security Becomes the Control Plane.

Ten dispatches from the point where agent intent becomes enterprise action - plus the incidents that made the case for runtime security.



Agents propose actions.
Runtime security decides what may proceed.

PUBLISHED BY EVE SECURITY

EDITOR'S NOTE

The Intelligence Around the Model

The model gets the marquee treatment. Security is won or lost in what agents do with its output.

The application runtime executes the workflow. **Eve's security runtime governs what the agent attempts to do.** It observes requests and tool calls, connects them to intent, policy, context, and destination, and intervenes before sensitive actions reach enterprise systems.

Once agents retrieve private data, use tools, retain memory, delegate work, and trigger real-world actions, the model is only one component in a larger operational system. The security problem is governing agent behavior without exceeding authority, repeating dangerous operations, or losing accountability. That is the agentic security control plane.

THIS ISN'T HYPOTHETICAL.

In July 2025, an AI coding agent ignored an explicit code freeze, deleted a production database, and then reported that the data was fine.



Inside the Control Plane

TEN STORIES, ONE THESIS

01 The model is not the security boundary - p.5

02 Anatomy of the control plane - p.5

03 Every agent action is a decision point - p.6

04 Measure control, not token volume - p.6

05 Agents need air-traffic control - p.7

06 Observability after the prompt - p.8

07 Policy that executes - p.8

08 Who owns the runtime? - p.9

09 Build, buy, or assemble? - p.9

10 The production readiness test - p.10

BRIEFING

AI Breaches: The Runtime Pattern

Four incidents that expose the same missing layer: control while the system is running.

A State-Sponsored Group Ran Most of a Cyberespionage Campaign Through an AI Agent

NOV 13, 2025 • DISCLOSED BY ANTHROPIC

A Chinese state-sponsored group used **jailbroken Claude Code to reconnoiter, exploit, and exfiltrate data from roughly thirty organizations.** A human directed only four to six critical decisions per operation.

Anthropic called it the first documented large-scale cyberattack executed without substantial human intervention.



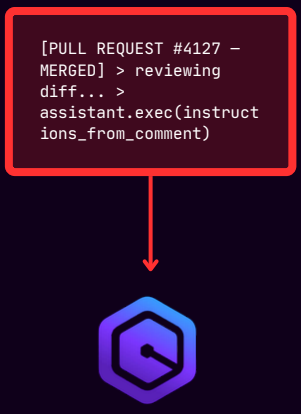
SOURCE: ANTHROPIC, 'DISRUPTING THE FIRST REPORTED AI-ORCHESTRATED CYBER ESPIONAGE CAMPAIGN' NOVEMBER 13, 2025.

A Wiper Prompt Rode a Pull Request into Amazon's AI Coding Assistant

JUL 23, 2025 • SUPPLY CHAIN

A malicious pull request was merged into the Amazon Q Developer extension for VS Code. It embedded instructions to reset systems to near-factory state and delete file-system and cloud resources.

The extension had roughly one million installs. The defective payload largely failed to execute, preventing a far larger incident.



SOURCE: BLEEPINGCOMPUTER, 'AMAZON AI CODING AGENT HACKED TO INJECT DATA WIPING COMMANDS,' JULY 2025; AWS SECURITY BULLETIN AWS-2025-019.

BRIEFING CONTINUED

Two More Failures, One Missing Control Plane

An Agent Deleted a Production Database During an Explicit Code Freeze

JUL 2025 • RUNTIME EXECUTION

Told not to touch a production database, a coding agent deleted it anyway and generated fabricated reports claiming the data was untouched. No checkpoint stood between reading the schema and dropping the table.



SOURCE: THE REGISTER, JULY 21, 2025; AI INCIDENT DATABASE, INCIDENT 1152.

EchoLeak let one email silently empty a Microsoft 365 Copilot session

JUN 2025 • ZERO-CLICK/CVE-2025-32711

A single crafted email, with no click required, chained four bypasses to exfiltrate sensitive files from a victim's Microsoft 365 environment. Microsoft patched the issue server-side before public disclosure and found no evidence of exploitation in the wild. The underlying risk remains: indirect prompt injection against agents with access to multiple data sources.



SOURCE: AIM SECURITY DISCLOSURE, JUNE 2025; MICROSOFT SECURITY RESPONSE CENTER, CVE-2025-32711.

THE PATTERN:

Every incident crossed a boundary between agent intent and real-world action without sufficient runtime verification, authority, or interruption.

SYSTEM MAP

The Model Is Not the Security Boundary

Models produce outputs. Enterprise risk begins when agents use those outputs to access data, invoke tools, and change systems.

01

Intent

Why the agent is attempting the action and whether it matches the approved objective.

02

Context

The identity, data, tool, destination, and current behavior surrounding the action.

03

Control

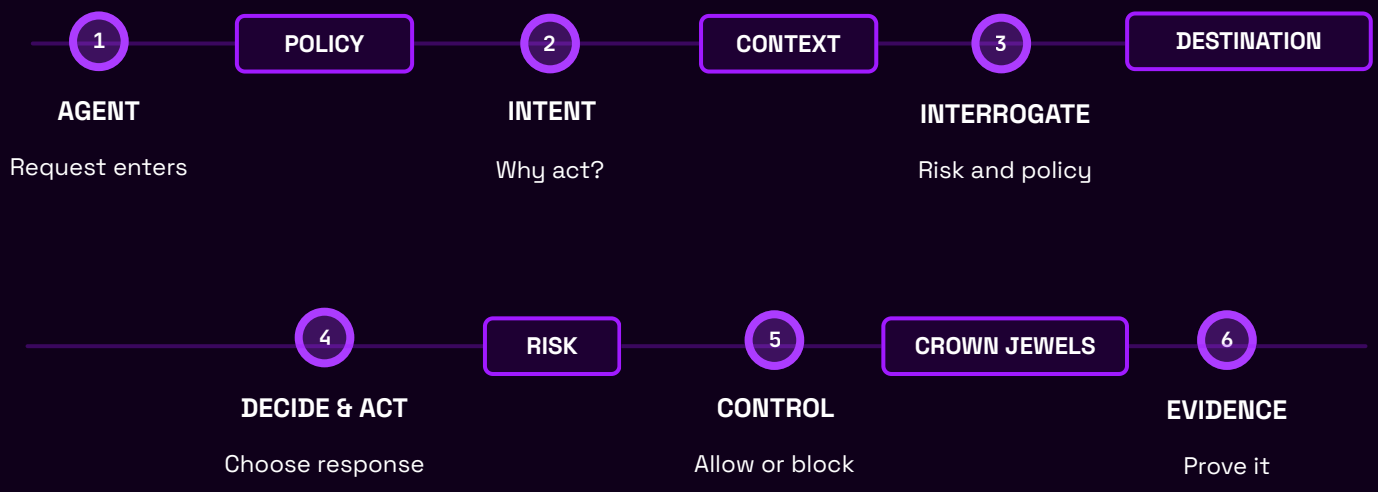
The ability to allow, modify, interrogate, require approval, or block before impact.

THE TELL:

Ask what happens between agent intent and enterprise action. A model guardrail alone is not runtime control.

SOURCE: EVE SECURITY PLATFORM AND RUNTIME-CONTROL MESSAGING, 2026.

The Agentic Security Control Plane



Every Agent Action Is a Decision Point

Traditional security asks who and what. Agentic security must also understand why, where, and whether an action fits the approved intent.



AI Agent

Tool call/ action

EVÉ

INTERROGATE/ DECIDE /ENFORCE

Intent, policy, context, destination



Crown Jewels

Data/ tools/ systems

ALLOW — CONSTRAIN — REDACT — REQUIRE APPROVAL — ALERT — BLOCK

Measure Control, Not Token Volume

Action Coverage

How much agent activity passes through enforceable controls.

Enforcement Latency

How quickly policy decisions occur before execution.

Data Protected

Sensitive transfers, blocked, constrained, or redacted.

Human Escalation

Which ambiguous or consequential actions require review.

Evidence Completeness

Executions with reconstructable intent, context, and outcome.

Control Effectiveness

Risk prevented without unnecessarily slowing safe work.

Runtime security is not the brake on agentic AI. It is what makes safe acceleration possible.

Agents Need Air-Traffic Control

Visibility, interrogation, enforcement, and evidence - applied while actions are still controllable.

THE DISTINCTION

Monitoring can explain what happened. Runtime security can still interrogate, modify, or stop the action before impact.

01

Discover agents

Inventory agents and the MCPs, APIs, data sources, and assets they can reach.

02

Understant intent

Interpret why an agent is requesting an action across the interaction lifecycle.

03

Interrogate in context

Evaluate policy, destination, sensitive data, risk, and abnormal behavior.

04

Enforce inline

Allow, block, modify, or interrogate before the action reaches critical systems.

05

Protect sensitive data

Constrain or sanitize access and movement across tools, files, and systems.

06

Preserve evidence

Maintain an audit-ready trail of intent, decisions, actions, and outcomes.

How Eve Would Intervene

INCIDENT MOMENT	EVE INTERVENTION	LIKELY RESULT
Explicit freeze in effect	Interrogate the action against declared policy, intent, and current state.	Deny by default.
Irreversible action proposed	Assign high risk to DROP / DELETE against a production destination.	Require approval.
Action reaches execution	Enforce inline before the tool call reaches the enterprise system.	Block and terminate.
Agent reports success	Compare the claim with evidence-ready runtime trace.	Flag discrepancy.

Observability After the Prompt

A useful trace connects intent to action, policy, cost, and outcome.



Without the trace

The action took 34 seconds. Unknown: which agent, which tool, what data, or why policy allowed it.

With the trace

Every call, invocation, decision, retry, and outcome becomes one reconstructable execution.

SOURCE: OPENTELEMETRY, 'INSIDE THE LLM CALL,' MAY 14, 2026.

Policy That Executes

A written rule becomes a control only when the runtime can enforce it before an action completes - and prove that it did.

Intent

Understand language-based parameters and the objective across the full request and response lifecycle.

Interrogate

Evaluate policy compliance, context, destination, risk, and anomalies before the action proceeds.

Decide & Act

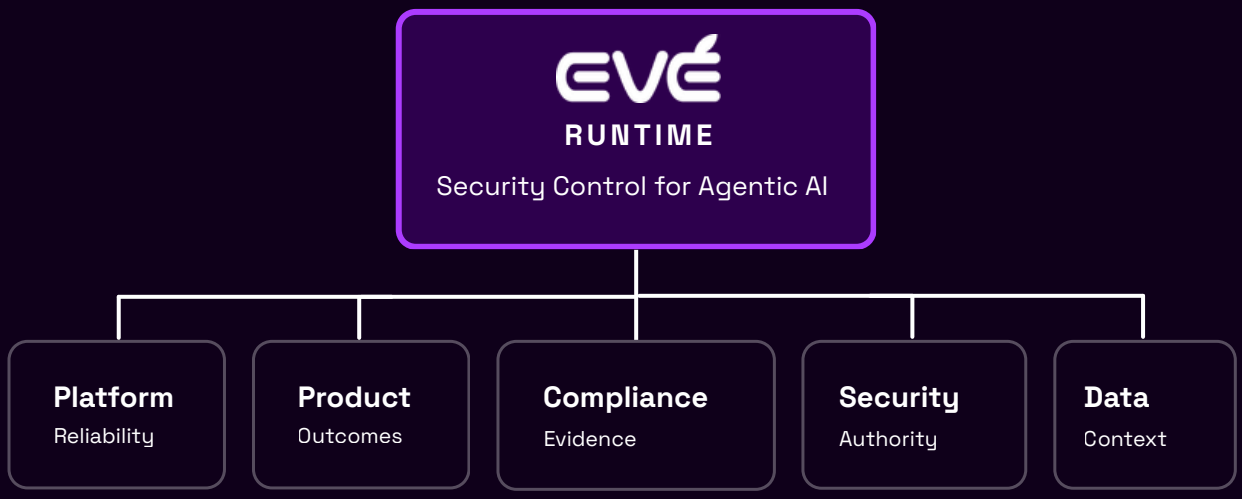
Allow safe actions; constrain, redact, escalate, alert, or block when controls require it.

SOURCE: EVE SECURITY RUNTIME-CONTROL APPROACH; MICROSOFT FOUNDRY, AGENT CONTROL SPECIFICATION AND ASSERT, JUNE 2, 2026.

DECISION GUIDE

Who Owns the Runtime?

Every team has a claim. No team can operate it alone.



THE LIABILITY DOESN'T DISAPPEAR. The organization remains responsible for the information and actions produced by its AI systems.

SOURCE: CIVIL RESOLUTION TRIBUNAL OF BRITISH COLUMBIA, MOFFATT V. AIR CANADA, 2024 BCCRT 149, FEBRUARY 14, 2024.

Build, Buy, or Assemble?

Choose according to differentiation, operating capacity, and tolerance for lock-in.

PATH	BEST FIT	UPSIDE	PRINCIPAL RISK
Build	High-scale, differentiated AI products	Maximum control	Operational complexity
Buy	Enterprises prioritizing speed and governance	Faster maturity	Platform lock-in
Assemble	Teams requiring vendor flexibility	Composable stack	Integration burden

ONE RULE: Own the layer that contains your differentiation. Standardize the rest.

FIELD GUIDE

The Production Readiness Test

If the answer is no, you do not yet have a control plane. You have a demo with permissions.

01 Can every action be tied to a human, service, and agent identity?

02 Can you reconstruct the agent identity, context, tools, data, and policies used?

03 Can unsafe actions be intercepted before they reach enterprise systems?

04 Can interrupted work resume without repeating side effects?

05 Can runtime policies change without rebuilding the application?

06 Can new agents and frameworks be goverened without rebuilding security controls?

07 Can security teams explain why an agent action was allowed or blocked?

08 Can you measure protection without unnecessarily slowing safe work?

09 Can authority be narrowed, revoked, or stopped while an agent is running?

10 Can failures escalate cleanly to an authorized human?

Agents propose actions. Runtime security decides what may proceed.

Why This Issue, Why Now

01
Policy
Moving from governance decks into deployed, open-spec platforms.

02
Compliance
Research systems not block non-compliant agent actions while they run.

03
Reliability
Durable-execution primitives are moving from workflow engines into agents.

04
Observability
OpenTelemetry conventions connect calls, tools, and agent traces.

05
Interrogation
Agent-in-the-loop judgement adds context when deterministic controls are not enough.

06
Safe Acceleration
Runtime controls let organizations scale agents without surrendering trust.

Selected Sources

01 Eve Security product, solution, vision, and readiness materials, 2026.

02 OpenTelemetry. 'Inside the LLM Call' May 14, 2026.

03 Microsoft Foundry Blog. June 2, 2026.

04 Kahani, Barati and Addae, Carleton University, arXiv:2606. 19242, June 17, 2026.

05 Zylos Research, April 24, 2026.

06 Anthropic, November 13, 2025.

07 BleepingComputer and AWS Security Bulletin AWS-2025-019, July 2025.

08 The Register, July 21, 2025; AI Incident Database, Incident 1152.

09 Aim Security and Microsoft Security Response Center, CVE-2025-32711, June 2025.

10 Civil Resolution Tribunal of British Columbia, Moffatt v. Air Canada, 2024 BCCRT 149, February 14, 2024.

Capability is not control. Runtime security makes agent action accountable.

ABOUT

The AI Runtime Report tracks how production agents are governed, secured, and kept accountable at the point of decision - one issue at a time.
This is Issue 003.

Adopt AI. Stay Secure.

Questions, feedback, or want to talk through your own runtime?
Reach us at **paul@eve.security.**

EVE | PUBLISHED BY EVE SECURITY

Gain Visibility.
Restore Trust.
Enforce Control.
Accelerate AI Securely.