SENSEON · SECURITY FOR AI

# Security changed shape when software started acting.

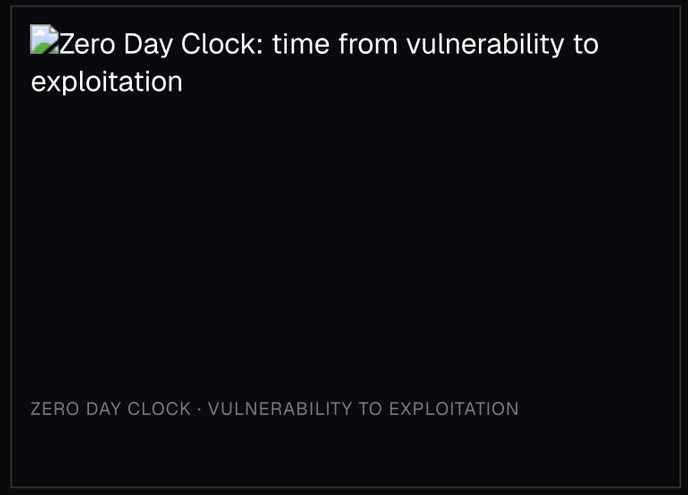
CISO DECISION GUIDE

For CISOs, security architects, and AI programme owners who must govern AI adoption without losing the ability to reconstruct what happened.

WHY NOW

# The exploit clock is collapsing while agent autonomy is lengthening.

Zero Day Clock public charts show exploitation tempo rising and the exploit window shrinking. Anthropic autonomy research shows agent sessions lengthening. Attackers are getting faster while agent decision chains get longer — the time you have to reconstruct and respond is compressing from both sides.


## Autonomy windows are already lengthening in practice.

Among Anthropic's longest Claude Code sessions, session length nearly doubled from under 25 minutes to over 45 minutes in three months. Longer autonomous runs mean longer decision chains to reconstruct.

Source: Anthropic, Measuring AI agent autonomy in practice, 18 February 2026 · cohort: longest sessions

OPERATING SHIFT

# Pre-AI controls govern access. Post-AI controls must govern decisions and actions.

PRE-AI MODEL

## Human uses software

- Identity authenticates a person
- Access policies gate systems and data
- Logs show who signed in and which tool was used
- Incident response reconstructs human-led sequences

POST-AI MODEL

## Agent reads, decides, and acts

- Agent identity, tools, and delegated privilege matter
- Content becomes an instruction-bearing attack surface
- Evidence must join model output, tool calls, and effects
- Response must interrupt or reverse agent actions in policy

CATEGORY

# Two foundational problems define the next security era.

They are not competing product lines. They are linked operating jobs joined by the same evidence architecture.

## AI for Security

Use governed agents to compress investigation and response work—with identity, tools, policy, approval, audit, and rollback under control.

## Security for AI

Observe and control AI providers, agents, identities, tools, data access, code, and the actions those systems take across the estate.

The bridge is an evidence architecture—not another provider dashboard. If you cannot reconstruct what an agent saw, decided, called, and changed, you cannot govern either job.



THREAT MODEL

# Agent risk is an identity, supply-chain, data, and action problem.

### Goal hijacking and tool misuse

Agents can be steered into unintended objectives and unsafe tool use. OWASP's Agentic Applications Top 10 places these at the centre of the new taxonomy.

### Identity and privilege abuse

Delegated credentials, service accounts, and broad tool scopes turn a compromised agent into a high-privilege operator.

### Agentic supply chain

Models, plugins, packages, knowledge sources, and agent templates introduce third-party risk into the decision path.

### Prompt injection as social engineering

OpenAI describes practical prompt-injection attacks as increasingly resembling social engineering against agents that browse, retrieve, and act.

### Unexpected code execution

Coding and operations agents can write, run, or merge code that alters production paths without a clean human-readable audit trail.

### Memory and context poisoning

Poisoned context and long-lived memory can bias later decisions across sessions and tools.

CISO PRIORITIES

# Five CISO decisions now need evidence.

Each decision fails if the organisation can only inspect provider consoles and disconnected logs.

01 **Adoption governance**

Which agents, providers, and tools are approved—and can you prove who used what under which policy?

02 **Supply chain**

Which models, packages, plugins, and knowledge sources sit on the path from prompt to production change?

03 **Identity and privilege**

What identities, tokens, and tool scopes can an agent exercise, and who can approve escalation?

04 **Data exposure**

What sensitive data entered a model context, left a controlled system, or moved through an AI workflow?

05 **Incident response**

Can responders reconstruct the full action chain and interrupt or reverse agent behaviour under policy?



PROVIDER EVIDENCE

# Provider controls are pieces of the puzzle, not the whole investigation.

Native logs are useful. They are not a complete security record across the estate.

| SOURCE | NATIVE PROTECTION | EVIDENCE CONTRIBUTED | GAP | SENSEON JOINS | DECISION ENABLED |
|---|---|---|---|---|---|
| **AWS Bedrock** | IAM, model access, guardrails, input/output checks | CloudTrail API events; optional invocation, agent, and knowledge-base logs | Other providers; device, network, owner, and downstream action | Invocation + identity + cloud + network + endpoint + code | Was the action expected, and what must be contained? |
| **Anthropic / Claude Enterprise** | SSO, SCIM, roles, retention controls | Enterprise user, system, and data-access audit events | Unmanaged use; tool effects; cross-estate lineage | Claude + identity + endpoint + GitHub + network + SaaS | Who used it, what was exposed, and what changed next? |
| **OpenAI / ChatGPT Enterprise** | Workspace identity, administration, API data controls | Compliance logs and state APIs; time-bounded unless exported | Personal accounts; other tools; downstream action; durable retention | Retained provider evidence + cross-estate context | Is this isolated AI use or part of a wider incident? |
| **Microsoft Copilot + Purview** | Tenant policy, identity, information protection, DSPM/DLP | User/admin audit, resources accessed, prompt/response content with DSPM | Other providers; endpoint/network behaviour; runtime outcome | Microsoft + external AI + device + network + code | Did protected data move, and was the action authorised? |

| | | | | | |
|---|---|---|---|---|---|
| **GitHub Copilot** | Enterprise policy, access, and agent controls | Usage, agent audit events, user, session ID, action | Local prompt/session; package/runtime behaviour; cloud/SaaS effects | Agent audit + repo + package + endpoint + identity + CI/CD | Did the agent introduce risky code, dependency, or credential use? |

Sources: official AWS Bedrock, Anthropic Compliance API, OpenAI Compliance API, Microsoft Purview, and GitHub Copilot documentation. Accessed 12 July 2026.



EVIDENCE CHAIN

# The decisive question is whether every action can be reconstructed.

A Security for AI programme is not complete until this chain is inspectable end to end.

01
### Input
Human prompt, retrieved content, or system trigger that starts the agent run.

02
### Source
Identity, device, repo, ticket, email, or data source that supplied context.

03
### Model output
What the model proposed, classified, or planned before tools ran.

04
### Tool call
Which tools, APIs, browsers, or systems the agent invoked, with parameters.

05
### Approval
Policy check, human gate, or automated allow/deny decision on the path.

06
### Action
What changed: code, ticket, mail, cloud resource, data, or access state.

07
### Correction
Containment, rollback, escalation, and the evidence retained for review.

ARCHITECTURE

# SenseOn connects the full chain: evidence, reasoning, agents, and governed action.

**AGENT CONTROL PLANE** Governance around every stage: identity, permissions, tools, policy, approval, audit, timeout, escalation, and rollback for every permitted action.

01

### Sources

Identity, endpoint, network, cloud, SaaS, AI providers, code, and business systems across the estate.

02

### Data Fabric

Ingest, shape, and retain source-linked evidence with lineage preserved and cost routed by value.

03

### Intelligence Fabric

Join entities, sequences, detections, and prior context so investigations stay source-linked as questions multiply.

04

### Horus · Orchestrator

Runs the case workflow and the permitted specialist agents operating on the evidence plane.

### Resolve · Investigator

Investigates evidence and completes or escalates eligible cases within defined response boundaries.

Decision enabled: which sources matter for this agent workflow?

Decision enabled: is the evidence retained and inspectable?

Decision enabled: what is connected across providers and tools?

Decision enabled: can this case close, or must a human take over?

Horus (orchestrator) and Resolve (Investigator) act only through the Agent Control Plane, which wraps the whole flow. Provider safeguards remain necessary; they do not replace cross-estate evidence.



DATA FABRIC

# Process and route evidence before central cost becomes a security constraint.

When ingestion economics force teams to drop sources or shorten retention, growth becomes a security problem—not only a platform cost problem.

### Edge processing

Fluent Bit mechanisms parse, filter, buffer, and route evidence close to where it is created, with memory and filesystem buffering that absorbs backpressure when destinations slow down. Noise with no security value never has to travel. Reduction is measured source by source.

### Source lineage

Preserve origin while normalising the common fields needed to search and join identity, endpoint, network, cloud, SaaS, AI-provider, code, and business-system evidence.

### Value-based retention

High-value evidence supports detection and response. Lower-value volume can remain searchable without occupying the most expensive path. Measure reduction source by source.

### ClickHouse evidence plane

Columnar storage, ordering keys, codecs, materialised fields, and sparse indexing let an investigation read the columns that answer its question instead of scanning every stored byte—so high-fidelity events stay retained and queryable. Query behaviour is measured per workload.



INTELLIGENCE FABRIC

# Fast search matters because agents ask more questions per investigation.

Query speed alone is not enough. Schema consistency, security semantics, API usability, and context determine whether the investigation holds.

### Fewer wasted bytes

### Repeated security semantics

Columnar access patterns and ordered evidence help investigations read the fields that answer the next question, not every stored byte.

Identity, host, process, flow, case, detection, and agent action must mean the same thing across sources—or joins invent false confidence.

**Source-linked reasoning**

Every conclusion stays attached to the originating evidence so responders can inspect, challenge, and export the basis of a decision.

**Operating work avoided**

The test is not only “how fast is the query?” It is how much manual stitching, console-hopping, and spreadsheet assembly the architecture removes.



FIELD LESSONS

# Three field lessons changed how we validate the architecture.

Each lesson is a test you can run on your own evidence — before you commit budget or rip out systems that already work.

LESSON 01 · COEXISTENCE

## Keep working platforms. Add the security evidence and action plane.

LESSON 02 · ECONOMICS

## Process, route, and retain evidence by value.

LESSON 03 · HANDS-ON PROOF

## Architecture slides are not validation.

Reliability, daily workflow, raw evidence behind a finding, custom-log handling, provider and agent visibility, false-positive control, and

A modern team may already run a lakehouse and agentic workflows. The winning route can be coexistence: preserve engineering systems that work, and prove reconstructability for security investigations and governed response.

Future data growth becomes a security problem when cost forces dropped sources or short retention. Compare suppliers on the same source coverage, retention, deployment burden, investigation workflow, and response outcome.

response boundaries belong on the scorecard. Prove them on real evidence.



PRIORITY USE CASES

# Start with the workflows that cross the most trust boundaries.

| | |
|---|---|
| **Agent and software supply chain** | Trace model, package, plugin, and repo paths that feed automated code or configuration change. |
| **Sensitive-data access through AI** | Join identity, DLP/DSPM, provider activity, and downstream action when protected content enters a model context. |

| | |
|---|---|
| **Coding agents** | Reconstruct who initiated a session, what was suggested or committed, and what ran in CI or production. |
| **AI for the SOC** | Govern investigation agents so case work accelerates without unscoped tool use or unaudited response actions. |
| **Provider sprawl** | Correlate Bedrock, Claude, OpenAI, Copilot, and GitHub activity with estate-wide identity, endpoint, and network evidence. |



VALIDATION PLAN

# Prove one chain of custody in 30 days.

One real agent workflow. The organisation’s own evidence. A pass/fail reconstruction test.

| WEEK 1 | WEEK 2 | WEEK 3 | WEEK 4 |
|---|---|---|---|

### Source map

Inventory AI providers, agent workflows, identities, tools, and the evidence each already produces.

Output: source coverage map and reconstruction gaps.

### Bounded workflow

Select one agent workflow that crosses multiple trust boundaries and define the chain-of-custody questions.

Output: workflow scope, pass/fail criteria, response boundary.

### Evidence pack

Join provider, identity, endpoint, network, code, and action evidence for the selected workflow.

Output: inspectable chain from input to correction path.

### Executive readout

Present risks found, remaining gaps, operating model, and recommended governed scope.

Output: decision pack for funding and next-step scope.



NEXT STEP

# Bring one agent workflow. Leave

# with an evidence-backed control plan.

## Request a Security for AI evidence assessment.

Scope: your sources, one agent workflow, reconstruction gaps, and a four-week validation path built around your operating environment.

The control plan is defensible when the organisation can prove ownership, data access, decision lineage, response, and control effectiveness on its own evidence.