Netzilo AI Detection & Response (AIDR)

The AI Control Plane for the Agentic Workforce

A brief for security leaders evaluating AI agent governance — CISOs, security directors, and the architects who will run it.

Your AI agents are the new BYOD.

A decade ago, unmanaged phones walked into the enterprise faster than security could govern them. Autonomous AI agents are doing it again — only now the workforce is silicon, it operates at machine speed, and it can read files, call tools, and reach the network on its own. The stack you built for human users cannot see what these agents intend to do.

⚠ 60 SECONDS, NO ALERT

A coding agent your team approved on Friday is compromised through a poisoned tool description. Here is the chain — and what each layer of your stack reports while it happens:

1. Agent ingests a malicious tool description (indirect prompt injection)
2. Acquires an external “skill” from an unknown host
3. Reads credentials and customer records from the local workspace
4. Sends them outbound to the attacker over a routine-looking API call

EDR SEES	SIEM SEES	AIDR SEES
Normal process telemetry. No malware.	A few benign-looking API calls.	**The whole chain — and stops it.**

That governance gap is the problem this document is about.

1 THE GAP YOUR STACK CANNOT CLOSE

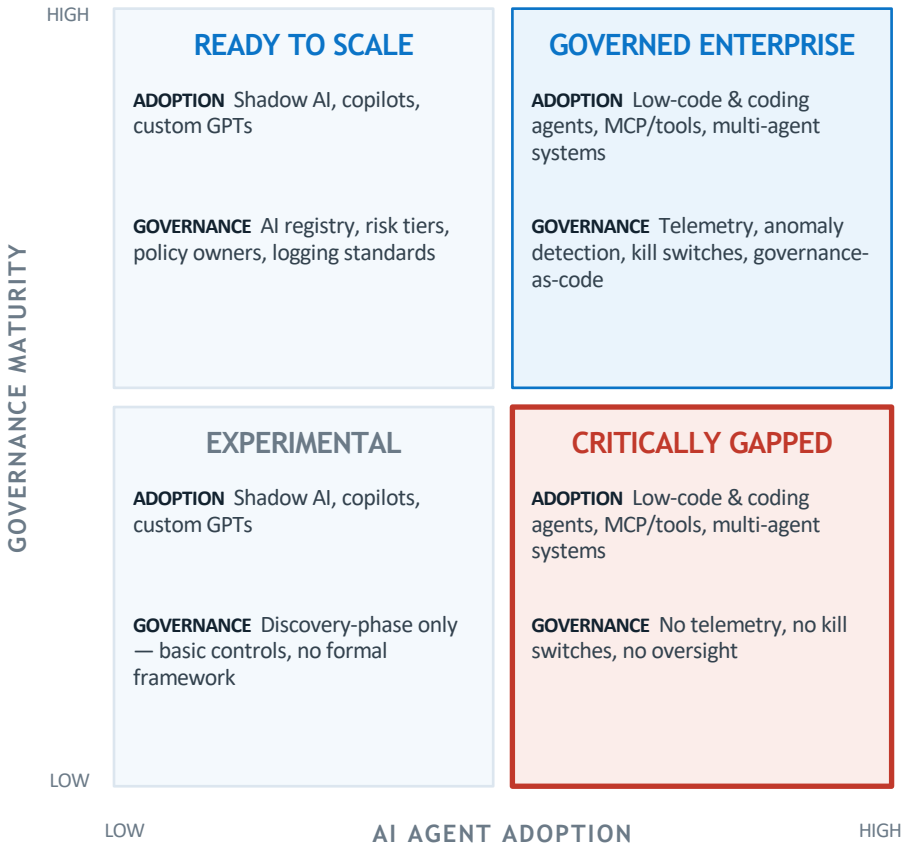
Built for humans, blind to agents

EDR and SIEM watch low-level telemetry — a file read, a process spawn, a network call. They have no semantic context for the objective behind those actions, so obfuscated payloads, prompt injection, and tool poisoning pass as ordinary activity.

Operating outside your gates

Agents talk over machine-speed protocols — Model Context Protocol (MCP) and Agent-to-Agent (A2A) — that never traverse the network chokepoints your controls sit on. The decision and the damage both land before a human could react.

Where most enterprises sit today



MOST ARE HERE

Roughly half of employees can already reach unsanctioned AI models and tools. Adoption has outrun governance — the bottom-right quadrant is the default state, not the exception.

THE ECONOMICS

Cost per attempt

ATTACKER
~$0.0002

STATIC GUARDRAIL
~$9.67

AIDR BEHAVIORAL
<$0.0001

Source: Elastic Security Labs · LLM Reversing vs LLM Obfuscation · 2026 | OWASP State of Agentic AI Security & Governance, June 2026

2 WHAT NETZILO AIDR IS

Not another language filter. A runtime control plane.

Guardrails supervise what a model says. AIDR governs what an agent does — at the level of tool calls, file access, and network activity, in real time. Three things you can do with it:

SEE

Discover the shadow AI you don't know about — unsanctioned LLM usage, unmanaged MCP servers, and off-book automation — with an endpoint-level activity graph and zero friction.

DETECT

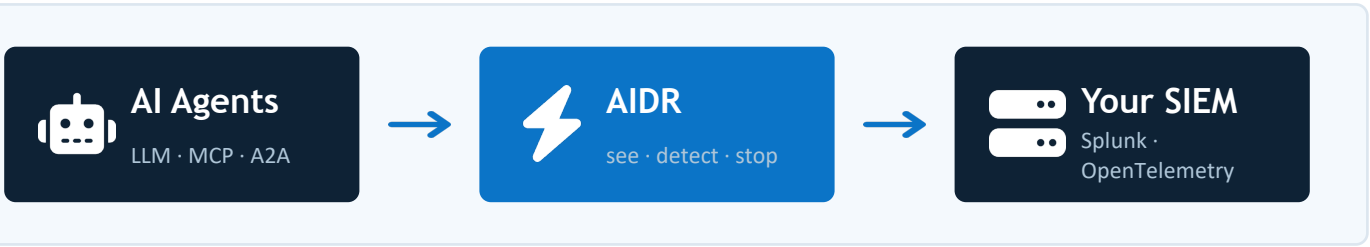
Catch what content filters miss: prompt injection, tool poisoning, privilege escalation, and multi-stage chains that look benign one event at a time but add up to a breach.

STOP

Enforce deterministic boundaries as code. When an agent crosses the line, the kill-switch isolates or terminates it instantly — no human in the millisecond loop.

How it fits the stack you already run

AIDR is not a rip-and-replace. It is the semantic layer between your agents and the tools you already operate — translating raw agent behavior into meaning, enforcing policy at the edge, and replaying enriched events upstream into your SIEM.



No enterprise data is routed through third-party infrastructure — AIDR runs cloud-delivered or self-hosted, with no capex and fast integration.

3 HOW IT WORKS

Behavioral intelligence, not pattern matching

The runtime graph. AIDR records every action every agent takes — tool calls, file reads, HTTP requests, LLM calls, process spawns, skill acquisitions — as a live, queryable graph. Threats live in sequences, and a graph can see sequences a regex never will.

Governance-as-Code. Policies are deterministic rules that query that graph at runtime. One rule catches the page-one breach as a single correlated chain:

```
skill acquired          from unknown host
then  file read         credentials / records
then  HTTP out          to that same host
→  VERDICT: BLOCK       chain correlated in <1s
```

Three engines, one verdict

Static Scanner
fast content checks — PII, secrets, known-bad strings

Behavior Scanner
graph-aware chain & rate detection

Kill-Switch
isolate or terminate on a true positive

Organizational context weighs agent roles, owners, and team-specific "normal" so a legitimate anomaly isn't treated like an attack.

Open detection rules. The rules aren't a black box — inspect, fork, and contribute community detections at **github.com/netzilo/aidr-sigma**

Three progressive deployment tiers

01 **Agent Plugin**
Embeds AIDR inside Claude, ChatGPT, and Gemini — monitors LLM & MCP at the agent layer.

02 **Browser Extension**
Extends coverage to browser-based AI and secures enterprise browsing across all major browsers.

03 **Netzilo Client**
Supersedes both — OS-level & arbitrary agent activity, plus Zero Trust access and workspace protection.

4 WHERE TO GO FROM HERE

Interested? Pick the door that fits your role.

This was a briefing, not a deep dive. Whichever seat you're in, here is the next, lighter-weight step — no procurement conversation required.

CISOs & DIRECTORS

Get the executive brief

- A board-ready governance assessment mapped to the OWASP maturity model
- Where your org sits today — and the gap to "governed enterprise"
- Compliance & audit framing for regulated environments

ARCHITECTS & ENGINEERS

Go straight to the tech

- Inspect the open detection rules on github.com/netzilo/aidr-sigma
- A technical deep-dive on the runtime graph and rule engine
- Start at the Agent Plugin tier — no infrastructure change to pilot

The agentic workforce is already inside the perimeter.

Governing it is no longer optional. Start seeing what your agents are actually doing.

Get Started → www.netzilo.com