Netzilo AIDR: The AI Control Plane for the Agentic Workforce

As organizations deploy autonomous AI agents, security initiatives must shift to governing run-time execution of AI agents. Netzilo AIDR (AI Detection & Response) provides the **deterministic runtime barrier** required to secure this **silicon-based workforce** operating at machine speed.

The Reality of the New Attack Surface

Traditional security stacks, including EDR and SIEM, are fundamentally blind to the intent of autonomous AI. While legacy tools monitor low-level process telemetry—such as file reads or execution—they lack the **semantic context** to understand the high-level objective behind these actions. This **"Context Gap"** leaves enterprises vulnerable to obfuscated payloads and novel techniques like **prompt Injection** or **tool poisoning** that appear benign to standard sensors.
Moreover, AI agents utilize machine-speed protocols such as Model Context Protocol (MCP) or Agent-to-Agent (A2A) that operate outside the visibility of traditional security gates.

The Maturity Deficit

Today, majority of enterprises are in **"Experimental or Critically Gapped"** state, where roughly half of employees could use unsanctioned AI models and tools. Hence the **Operational Risk** of unmonitored automation has outpaced traditional governance.

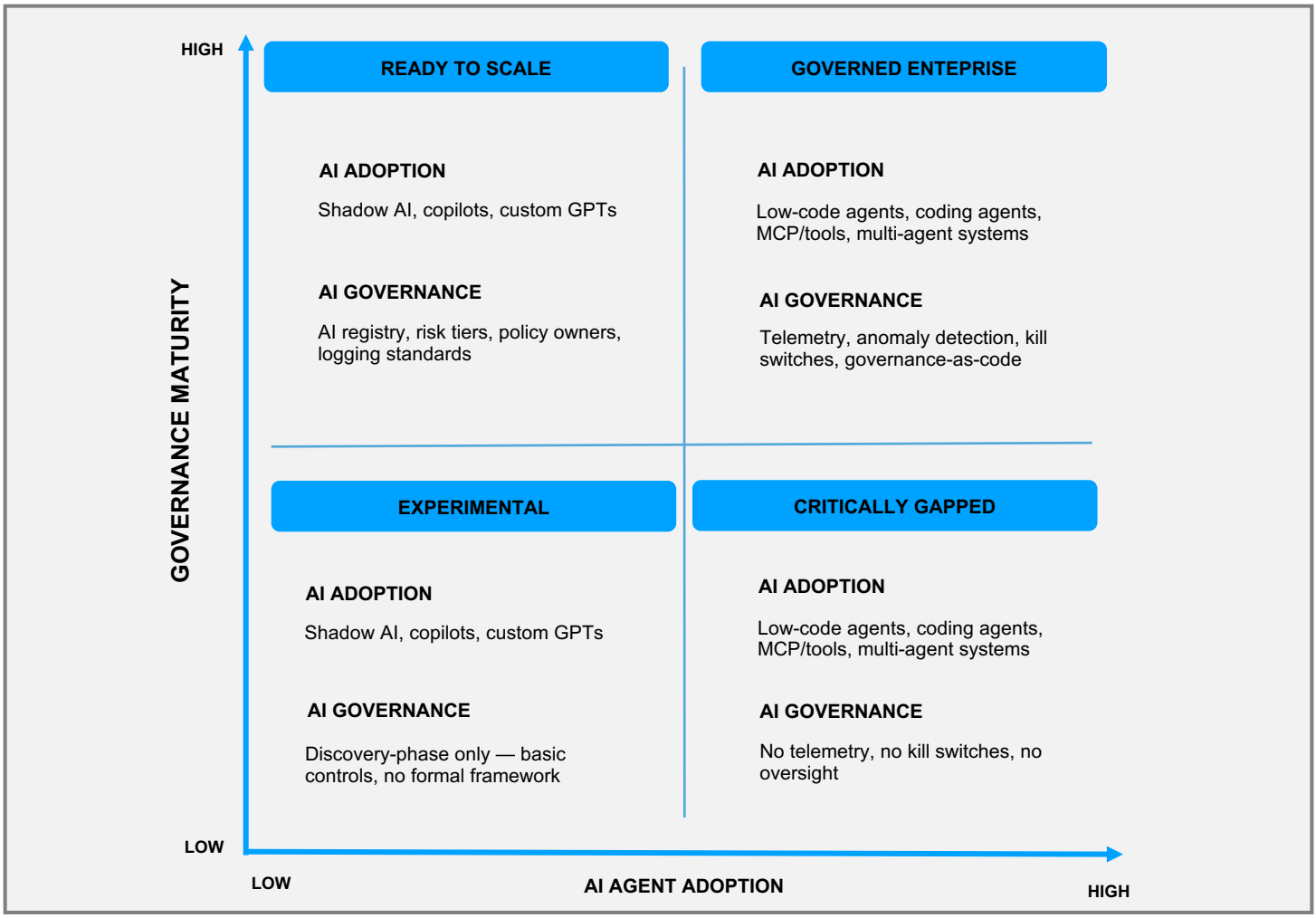


DIAGRAM 1: OWASP AI Agent Governance Maturity Matrix

AI Agent Governance

Netzilo AIDR delivers **immediate operational compliance** as a cloud-delivered service with **no capex** and fast integration into existing workflows. Its **Guardian Agent** provides endpoint-level activity context with zero friction, helping organizations discover unauthorized LLM usage, unmanaged MCP servers, and other "off-book" AI automation. With **Governance-as-Code**, Netzilo enforces deterministic controls, logging, and auditability for regulated industries.

According to the *OWASP State of Agentic AI Security and Governance (June 2026),* the transition to autonomous agents requires a fundamental shift in defensive architecture.

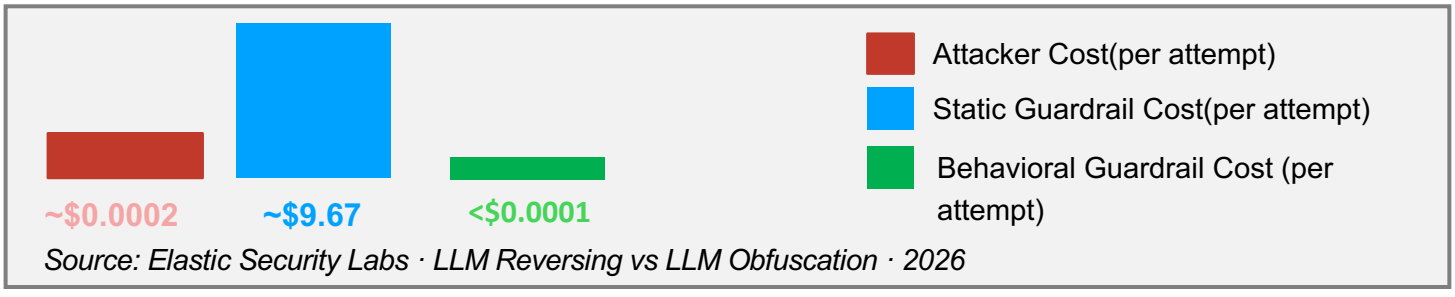
> "As AI agents' autonomy level increase, AI governance requires, deterministic enforcement, continuous behavioral monitoring, and kill-switch capability**."**
>
> *Source: OWASP · State of Agentic AI Security and Governance, June 2026*

As agent autonomy increases, enterprise governance demands behavioral intelligence and enforcement. While standard **"AI Guardrails"** focus on supervising language for LLM safety, they cannot govern apps, data, or runtime behavior.

Since traditional security is blind to novel techniques, Netzilo AIDR provides **cross-platform, multi-agent behavioral intelligence:**

- **Full Agent Visibility:** Rather than inspecting language only (which can be manipulated), Netzilo builds and inspects the **runtime graph**, including process execution, file system interactions, semantic actions such as tool calls, skill acquisitions etc. and network connections.
- **Multi-Stage Chain Correlation:** Identifies complex attack patterns by correlating sequences of actions that appear benign in isolation but indicate a breach in aggregate.
- **Organizational Context:** Leverages **Organizational Intelligence** to distinguish between a legitimate anomaly and an attack by factoring in agent roles, owners, and team-specific "normal" patterns

The Netzilo approach rebalances **the Economic Ratio to make defense cheaper** than the attack.



Source: Elastic Security Labs · LLM Reversing vs LLM Obfuscation · 2026

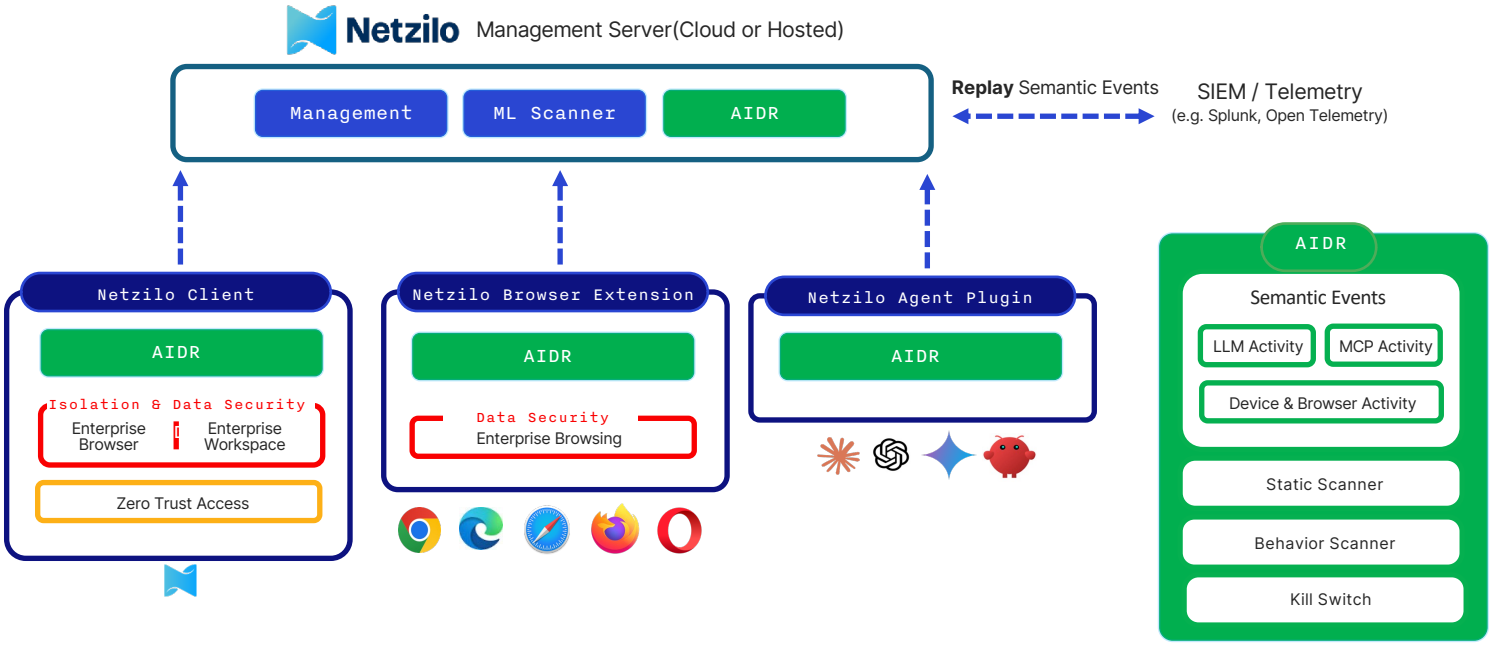


DIAGRAM 2: Netzilo Solution Architecture

The platform is built on four tightly integrated components that together deliver end-to-end AI Agent governance and security solution: the **Netzilo Management Server**, **Netzilo Client**, **Netzilo Browser Extension**, and **Netzilo AI Agent Plugin**.

The **AIDR (AI Detection and Response)** is the platform's core security engine, purpose-built to monitor, detect, and respond to AI-driven threats in real time. It captures semantic events across three dimensions — LLM activity, MCP communications, and device and browser behavior — feeding them through a Static Scanner, Behavior Scanner, and Kill-Switch to detect anomalous agent actions, privilege escalation, and prompt injection attacks the moment they occur. Unlike perimeter-based approaches, **AIDR treats every AI agent as a machine identity** subject to continuous behavioral scrutiny, **enforcing privilege boundaries and isolating or terminating compromised agents instantly** without routing any enterprise data through third-party infrastructure.

The AIDR is deployed through three progressive methods, each extending coverage deeper into the enterprise environment.

The **AI Agent Plugin** provides the entry-level deployment, embedding AIDR directly within AI platforms such as Claude, ChatGPT, and Gemini to monitor LLM and MCP activity at the agent layer. The **Browser Extension** provides these capabilities for browser-based flows, additionally securing enterprise browsing sessions across all major browsers, capturing a broader behavioral context including web-based AI interactions. The **Netzilo Client** supersedes both by extending AIDR to operating system activity and arbitrary AI agent activity, with additional capabilities such as Zero Trust access and enterprise workspace protection.