WHITEPAPER

MANAGED AI DETECTION & RESPONSE:

Securing the Enterprise AI Lifecycle with Vijilan Security

A comprehensive guide for MSPs, MSSPs, and enterprise security leaders on securing AI-powered workflows through a fully managed AIDR service.

99%	<30ms	50+	24/7
Prompt injection detection efficacy	Inline inspection latency	PII & PHI types detected	Human SOC coverage

EXECUTIVE SUMMARY

The AI Attack Surface Is Open. Most Organizations Have No Visibility Into It.

Enterprises are deploying generative artificial intelligence at unprecedented scale. 80% of employees now use AI tools at work — most without IT approval, many sharing sensitive data with public models without realizing the risk. At the same time, adversaries are weaponizing the same AI technologies: AI-enabled attacks increased 89% year-over-year according to the CrowdStrike 2026 Global Threat Report, and prompt injection now ranks as the #1 vulnerability for LLM applications on the OWASP Top 10.

Traditional security tools — endpoint detection and response (EDR), SIEM, identity platforms — were designed to protect infrastructure. They cannot see inside a prompt. They cannot detect when an employee pastes trade secrets into ChatGPT, when an adversary injects hidden instructions into a corporate AI agent, or when a RAG system exposes confidential documents to an unauthorized user. A new security category is required: AI Detection and Response (AIDR).

Vijilan Security delivers the industry's first fully managed AIDR service — powered by the same enterprise-grade AI security platform trusted by 60% of the Fortune 500. This white paper explains the threat landscape driving AIDR adoption, how the technology works, how Vijilan operationalizes it as a managed service, and why now is the critical window for MSPs, MSSPs, and enterprises to act.

> **KEY FINDING**
>
> IBM's 2025 Cost of Data Breach Report found that organizations breached via Shadow AI paid an average of $670,000 more per incident than standard breaches — and 63% of organizations still have no AI governance policy in place.

SECTION 1

The AI Threat Landscape: Why Traditional Security Falls Short

The Rise of Shadow AI

Shadow AI — the unsanctioned use of AI tools and models by employees without IT knowledge or approval — has become the fastest-growing unmanaged risk in enterprise security. According to Menlo Security's 2025 telemetry, enterprise visits to AI websites reached 10.53 billion in January 2025 alone, a 68% surge year-over-year. Only 22% of employees use AI tools exclusively approved by their employer.

The Shadow AI Problem

- 80% of employees use AI tools at work
- 78% bring their own AI tools to the workplace
- 57% actively hide their AI usage from employers
- 63% of organizations lack AI governance policies
- Only 17% have technical controls to prevent sensitive data uploads
- 86% of organizations are blind to AI data flows

The Financial Consequences

- $670,000 average breach cost premium for Shadow AI incidents
- 65% of breached data involved customer PII
- 40% involved intellectual property
- 20% of organizations experienced a breach attributable to Shadow AI
- Data shared with AI apps increased 30x in one year
- Average organization shares 7.7GB per month with AI services

The Prompt Injection Epidemic

Prompt injection has emerged as the defining AI-era security threat. It tops the OWASP Top 10 for LLM Applications 2025 and represents a fundamentally new class of attack that exploits the cognitive layer of enterprise AI rather than its infrastructure. Attackers craft malicious inputs — embedded in user messages, documents, web pages, or data sources — that manipulate AI models into performing unauthorized actions, bypassing safety constraints, or leaking sensitive information.

Real-world exploits illustrate the severity: CVE-2025-32711 ("EchoLeak"), a CVSS 9.3 vulnerability in Microsoft Copilot, demonstrated how hidden prompt injection in a PowerPoint presentation's speaker notes could cause Copilot to return a user's private emails to an attacker. Google paid $350,000 in AI-specific bug bounties in 2025, many linked to prompt injection vulnerabilities.

> **CROWDSTRIKE 2026 GLOBAL THREAT REPORT**
>
> "Prompts are the new malware." Adversaries exploited generative AI tools at 90+ organizations via malicious prompt injection in 2025. LLM-crafted phishing campaigns achieved a 54% click-through rate — more than four times the rate of human-generated attacks.

The Agentic AI Frontier

As enterprises deploy autonomous AI agents — systems that make decisions and take actions without direct human oversight — the attack surface expands dramatically. Gartner predicts 33% of enterprise applications will include agentic AI by 2028 (up from under 1% in 2024). The first documented AI-orchestrated cyberattack occurred in September 2025, when a Chinese state-sponsored group manipulated an AI coding agent to infiltrate approximately 30 global targets.

Non-human identities (NHIs) — AI agents, service accounts, API keys — now outnumber human identities 50:1 in enterprise environments. Traditional identity security tools were not designed to govern or monitor them. Autonomous agents can be manipulated via indirect prompt injection to expose sensitive credentials, install command-and-control malware, or exfiltrate data at machine speed.

SECTION 2

What Is AI Detection and Response?

AI Detection and Response (AIDR) applies the principles of Endpoint Detection and Response (EDR) to the interaction layer of enterprise AI — the prompts users send, the responses AI models return, and the communications between autonomous agents. Just as EDR monitors and responds to suspicious process behavior at the endpoint, AIDR monitors and responds to suspicious or malicious behavior within AI workflows.

The industry's leading AIDR platform, launched at Fal.Con 2025 and reaching general availability on December 15, 2025, unifies AI interaction security with the broader Falcon security platform. It introduces a new model: one platform for endpoint, cloud, identity, and AI security — one console, one investigation workflow, one source of cross-domain truth.

The Four Core Functions of AIDR

1. Discover Continuously inventory every AI application, agent, LLM runtime, MCP server, IDE extension, and cloud AI service across the enterprise environment. Know what is running before adversaries exploit it.	**2. Detect** Identify prompt injection attempts, jailbreak attacks, sensitive data leakage, malicious entity references, and unauthorized AI usage in real time — at sub-30ms latency with 99% detection efficacy.
3. Govern Enforce AI usage policies at the interaction layer. Redact sensitive content before it reaches AI models. Block unauthorized AI applications. Generate immutable audit logs for compliance frameworks.	**4. Respond** Take automated action — log, redact, or block — and surface AI security findings in the SOC alongside endpoint and identity telemetry for unified analyst investigation and response.

The Six Collector Architecture

AIDR captures AI telemetry through six purpose-built collector types, providing comprehensive visibility across every surface where AI interacts with the enterprise environment:

COLLECTOR	DEPLOYMENT	TELEMETRY CAPTURED
Browser	Chrome, Edge, Firefox extensions	Employee interactions with ChatGPT, Claude, Gemini, Copilot, and other public AI tools in managed browsers
Application	SDK (Python, Node.js, Go, Java, C#) + OpenTelemetry	Prompts and responses within enterprise-built AI applications at the runtime level
Agentic	MCP Proxy (Model Context Protocol)	AI agent-to-agent communications, tool calls, and autonomous decision-making across MCP clients and servers
Gateway	Kong, LiteLLM, Portkey, Apigee, Azure API Mgmt	AI traffic at API gateway boundaries without application code changes
Cloud	AWS Bedrock via S3 log ingestion	AI-related log events from cloud AI infrastructure and backend AI services
Endpoint	Falcon agent integration (pre-beta, GA Q2 2026)	Desktop AI applications: ChatGPT, Copilot, GitHub Copilot, Cursor, DeepSeek

SECTION 3

The Vijilan Managed AIDR Service

Vijilan Security is an authorized CrowdStrike Powered Service Provider (CPSP) delivering AIDR as a fully managed service through our global 24/7 Security Operations Center. Where most providers offer AIDR as a platform to be self-managed, Vijilan operates it — handling deployment, configuration, monitoring, threat response, and ongoing governance — so your organization receives outcomes, not tooling.

What We Deliver

Shadow AI Discovery & Governance
Continuous inventory of every AI application, agent, LLM runtime, MCP server, and cloud AI service across managed endpoints and cloud environments. Real-time policy enforcement — log, redact, or block — applied at the point of AI interaction.

AI Data Leakage Prevention
Automated detection and redaction of 50+ types of PII, PHI, financial data, credentials, proprietary source code, and trade secrets before they reach AI models. Format-preserving encryption available for structured sensitive data.

Agentic AI Monitoring & Control
MCP Proxy deployment to monitor and control all AI agent-to-agent communications, tool calls, and autonomous decision-making. Detection of rogue agent behavior and manipulated agent workflows in real time.

Prompt Injection Defense
Real-time detection and blocking of direct and indirect prompt injection, jailbreak attempts, and adversarial instructions using layered defense-in-depth: heuristics, neural classifiers, and dedicated GPU/CPU analyzers. 99% efficacy across 180+ tracked injection techniques.

24/7 SOC Integration & Response
AIDR telemetry streams into Falcon Next-Gen SIEM alongside endpoint, identity, and cloud signals. Vijilan's human SOC analysts triage AI-specific alerts with full cross-domain context and median critical response under 15 minutes.

Compliance Evidence Generation
Immutable, cryptographically signed AI interaction logs. Monthly AI risk reporting, policy tuning, and compliance evidence packages mapped to HIPAA, PCI-DSS, SOC 2, GDPR, NIST AI RMF, and EU AI Act requirements.

The Managed Service Delivery Model

Vijilan operates AIDR through a structured engagement that begins with comprehensive discovery and delivers continuous governance:

- Phase 1 — Discovery & Design: AI surface mapping, collector architecture design, policy framework development, compliance requirement alignment
- Phase 2 — Deployment: Browser extension rollout, SDK integration, gateway connectors, cloud ingestion pipelines — minimal disruption, sub-30ms latency impact
- Phase 3 — SOC Activation: AIDR telemetry integration into Vijilan's Next-Gen SIEM, alert correlation rules, analyst playbook development
- Phase 4 — Continuous Operations: 24/7 monitoring, monthly AI risk reporting, quarterly policy tuning, executive briefings, compliance evidence packages

Three Deployment Models

Vijilan supports all three AIDR deployment architectures to accommodate varying compliance, data residency, and performance requirements:

MODEL	MANAGEMENT	DATA PROCESSING	BEST FOR
SaaS	Fully Hosted	Vijilan/CrowdStrike cloud	Rapid deployment, minimal overhead, maximum simplicity
Edge	Hybrid	Customer AWS/Azure/GCP	Data residency requirements, latency-sensitive workloads
Private Cloud	Self-Managed	Customer infrastructure	Maximum sovereignty, defense & intelligence sectors

SECTION 4

Compliance Coverage: Every Framework, Every Requirement

Managed AIDR is purpose-built to support compliance evidence generation across the frameworks that govern AI usage in regulated industries. Vijilan operationalizes each framework mapping through specific technical controls, audit artifacts, and reporting deliverables.

HIPAA: Protecting PHI in Healthcare AI

Healthcare organizations face an acute dilemma: AI accelerates clinical and administrative operations, but HIPAA regulations impose strict controls on Protected Health Information (PHI). HHS OCR's proposed HIPAA Security Rule update (January 2025) — the first major revision in 20 years — removes the distinction between required and addressable standards, making AI-specific security controls effectively mandatory.
Managed AIDR addresses this through: real-time PHI detection and redaction using 250+ classification rules before data reaches AI models; immutable, cryptographically signed audit logs of all AI interactions involving clinical data; and BAA-compliant deployment architectures that satisfy HIPAA Business Associate Agreement requirements.

PCI-DSS v4: Financial Data in AI Contexts

Payment card data is increasingly processed alongside AI workflows for fraud detection, customer service automation, and financial modeling. PCI-DSS v4 Requirement 5, validated by Coalfire (a PCI QSA), is fully supported through AIDR's format-preserving encryption (FPE) capability — encrypting card numbers into same-format strings so AI can recognize data structure without accessing raw cardholder values.

SOC 2 Type II: Mapping to All Five Trust Service Criteria

AIDR maps directly to all five Trust Service Criteria: Security (blocks unauthorized AI access, real-time injection detection); Availability (monitors AI system health, detects disruptive attacks); Processing Integrity (detects model manipulation and data poisoning); Confidentiality (redacts credentials, PII, and trade secrets); and Privacy (shadow AI discovery, data leakage prevention). Immutable logs with cryptographic signing directly satisfy SOC 2 audit trail requirements.

NIST AI RMF: All Four Functions Operationalized

• Govern: AI usage policies, governance dashboards, accountability frameworks • Map: Shadow AI discovery, automated AI asset inventory, risk classification	• Measure: Continuous monitoring, 99% detection benchmarks, monthly risk reporting • Manage: Real-time blocking, automated redaction, SIEM integration, incident response

EU AI Act: Preparing for €35M Penalties

The EU AI Act imposes penalties up to €35 million or 7% of global turnover for non-compliance. Managed AIDR supports risk classification through AI inventory, technical documentation through audit logging, data governance through redaction, and human oversight through analyst-reviewed log/redact/block decisions.

SECTION 5

The Market Opportunity: Why Now

Three structural forces converge to make managed AIDR a high-urgency, high-growth opportunity for MSPs, MSSPs, and enterprise security teams:

Force 1: AI Adoption Is Outpacing AI Security

$37 billion in enterprise GenAI spending in 2025 — a 3.2x year-over-year increase (Menlo Ventures). 88% of organizations are using AI in at least one function. Yet 63% have no AI governance policy. The investment in AI capability is racing ahead of the investment in AI security. This gap is the opportunity.

Force 2: The Cybersecurity Talent Shortage Is Acute

4.8 million unfilled cybersecurity positions globally (ISC2 2025). 88% of organizations experienced at least one significant cybersecurity event due to skills shortages. AI is now a top-5 in-demand cybersecurity skill, yet most enterprises cannot hire fast enough. 94% of organizations say they are willing to pay more for AI security services from managed providers, with 70% prepared to pay 10-25% more.

Force 3: Gartner Has Declared AI Security a Top Strategic Priority

Gartner named AI Security Platforms a top strategic technology trend for 2026 and predicts that by 2028, more than 50% of enterprises will use AI security platforms (up from under 10% today). The firm also predicts 50% of all enterprise cybersecurity incident response efforts will focus on AI-related incidents by 2028.

CHANNEL MOMENTUM

CrowdStrike's MSSP business grew from sub-$100M to $1.3 billion in under three years. Vijilan is positioned as a premium channel delivery partner for the AIDR service expansion — bringing managed SOC, compliance expertise, and partner delivery infrastructure to a category that is growing faster than any other segment in cybersecurity.

SECTION 6

Why Vijilan: The Managed Provider Built for This Moment

Three structural forces converge to make managed AIDR a high-urgency, high-growth opportunity for MSPs, MSSPs, and enterprise security teams:

CrowdStrike Powered Service Provider
Vijilan is an authorized CPSP delivering AIDR natively on the Falcon platform — not a reseller, not an integration layer. We operate the same enterprise-grade technology trusted by 60% of the Fortune 500, through our own global SOC infrastructure.

24/7 Global SOC
Human analysts across global SOC locations. AIDR alerts are triaged alongside endpoint, identity, and cloud events — giving complete cross-domain context. One platform, one investigation, one response team.

Dual-Certified Security Controls
SOC 2 Type II and ISO 27001 certified by independent auditors. Vijilan holds the same standards we help clients enforce — independently verified, continuously maintained.

Platform Certifications
- SOC 2 Type II (A-LIGN audited)
- ISO 27001:2022 certified
- CrowdStrike Powered Service Provider (CPSP)
- FedRAMP High (Charlotte AI / Falcon platform)
- PCI-DSS v4 (Coalfire validated)
- HIPAA (Coalfire verified)

Performance Benchmarks
- 99% prompt injection detection efficacy
- Sub-30ms inline inspection latency
- 50+ PII/PHI types detected and redacted
- 180+ injection techniques tracked
- 26 programming languages for code detection
- 8 of 10 OWASP LLM Top 10 risks covered

Ready to Secure Your AI Attack Surface?

Schedule a consultation with Vijilan Security to design a Managed AIDR deployment tailored to your environment, your clients, and your compliance requirements.

vijilan.com/msp-partner-program | SOC 2 Type II | ISO 27001| CrowdStrike CPSP