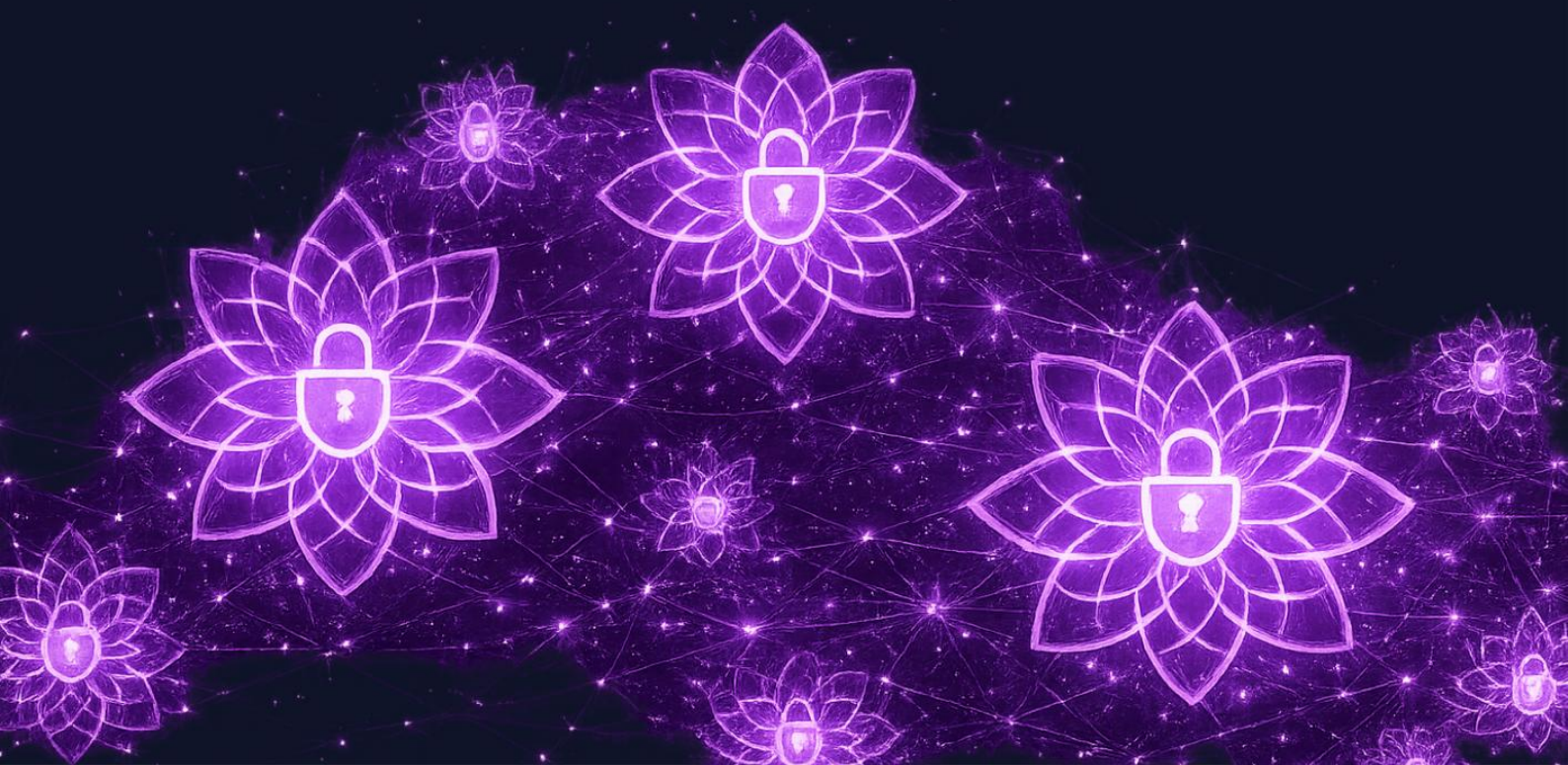
FIOR

Authentication for the Intelligent Age

Securing Agentic AI in the Enterprise

A Control-Plane Security Architecture for Mitigating MITRE ATLAS OpenClaw-Class Threats

Prepared for: Enterprise Chief Information Security Officers
Author: FIOR Group
Date published: February 2026



Executive Summary

Autonomous agentic AI systems represent a structural shift in enterprise risk. Unlike traditional applications, agentic systems ingest untrusted content, reason probabilistically, invoke tools, modify configuration and persist memory - often within a single trust domain.

This collapse of separation between ingestion and execution creates systemic exposure that cannot be addressed solely through model guardrails or conventional network controls.

The MITRE ATLAS OpenClaw Investigation (2026) provides a public, structured analysis of these risks. Documented attack chains included prompt injection leading to command-and-control persistence, malicious skill supply chain compromise, exposed control-plane takeover, privilege escalation, credential harvesting and autonomous data exfiltration.

These were not isolated bugs; they were architectural consequences of collapsed trust boundaries.

For enterprise CISOs, the strategic implication is clear: autonomous AI must be governed by an independent security control plane that mediates every execution request, enforces cryptographic identity, segments trust domains and maintains continuous auditability.

FIOR.AI Gateway introduces such a control plane

Positioned between AI reasoning layers and enterprise execution environments, it restores deterministic policy enforcement, enforces least privilege at tool level, segregates memory trust tiers and introduces accountable autonomy via human-in-the-loop controls.

The Enterprise Risk Landscape for Agentic AI

Enterprise adoption of AI has moved from passive inference systems to active digital operators. These operators can execute code, manage infrastructure, access data repositories and interact with production systems.

This fundamentally changes the security equation. Traditional controls assume:

- Human initiation of privileged actions
- Deterministic code paths
- Segregation between input and execution layers
- Static identity boundaries

Agentic systems violate each of these assumptions.

CISOs must therefore treat agentic AI not as a software feature but as a new operational identity class requiring dedicated governance.

MITRE ATLAS OpenClaw Threat Deconstruction

MITRE identified several recurring technique classes across OpenClaw deployments. The significance lies not only in individual exploits but in their chaining potential.

Prompt Injection and Indirect Control

Untrusted content embedded instructions that triggered tool invocation, allowing silent execution without policy validation.

Malicious Skill Supply Chain

Third-party skills executed with full privileges, lacking sandboxing or signature enforcement.

Control Plane Exposure

Internet-facing interfaces enabled credential harvesting and configuration modification.

Context Poisoning

Undifferentiated memory allowed low-trust input to influence high-privilege actions.

Credential Harvesting & Exfiltration

Static credentials and unrestricted tool invocation enabled silent data loss.

Architectural Root Cause: Collapsed Trust Boundaries

Across all documented incidents, ingestion, reasoning, execution and memory resided within a single privilege domain. This created a scenario where untrusted inputs could directly influence destructive actions.

CISOs must recognise that guardrails embedded inside the model are insufficient. Security must operate independently of model reasoning.

FIOR.AI Gateway Control Plane Architecture

FIOR.AI Gateway introduces a deterministic enforcement layer between AI reasoning and enterprise systems.

Core architectural components:

- Cryptographic workload identity
- Policy decision engine external to the model
- Tool-level capability contracts
- Memory trust segmentation
- Immutable telemetry logging

MITRE ATLAS Technique Crosswalk

MITRE Technique	Risk Description	Enterprise Impact	FIOR.AI Gateway Mitigation
Prompt Injection	Unauthorised tool invocation	Operational compromise	Policy-gated tool execution
AI Supply Chain Compromise	Malicious skill execution	Persistent implant risk	Skill attestation and sandboxing
Modify Agent Configuration	Security disablement	Control-plane bypass	Configuration lock and approval workflow
Credential Harvesting	Secret exposure	Lateral movement	Ephemeral, scoped credentials
Context Poisoning	Persistent memory corruption	Silent C2 channel	Trust-tiered memory segmentation
Exfiltration via Tools	Unauthorised data transmission	Regulatory breach	Egress policy and HITL approval

Governance and Accountable Autonomy

High-impact operations can require explicit human authorisation. This enables accountable autonomy without eliminating efficiency gains. All execution events are logged in tamper-evident audit trails, supporting forensic review and board-level reporting.

Regulatory & Assurance Alignment

FIOR.AI Gateway supports emerging compliance frameworks including:

- EU Cyber Resilience Act (CRA)
- NIS2 Directive
- Sovereign cloud mandates
- Sector-specific operational resilience standards

By separating policy enforcement from probabilistic reasoning, FIOR provides continuous assurance rather than reactive monitoring.

The Agentic AI Control Plane Category

The enterprise AI security market is converging toward a new category: **Control Planes for Autonomous Digital Operators.**

FIOR.AI Gateway is positioned distinct from:

- Model-level guardrail providers
- Traditional firewalls
- Endpoint detection platforms

It governs autonomy rather than traffic or endpoints.

5 Strategic Recommendations for CISOs

CISOs deploying agentic AI should:

1. Prohibit direct execution from model outputs
2. Enforce cryptographic identity for agents
3. Segment ingestion and execution domains
4. Require policy validation external to the model
5. Implement immutable audit logging

Conclusion

Agentic AI introduces a qualitatively new risk class. MITRE's OpenClaw investigation demonstrates that the threat is architectural rather than incidental.

FIOR.AI Gateway restores structural trust boundaries, enabling enterprises to deploy autonomous systems with governance, assurance and control.