Operation OpenClaw: Modeling an Agentic Kill Chain Against Enterprise Infrastructure

Fabrice Pizzi — Université Paris Sorbonne, 2026

Academic Summary Note — February 2026

Abstract

This paper presents a comprehensive threat model of a multi-phase cyberattack exploiting an autonomous AI coding agent (OpenClaw) as both attack vector and force multiplier against a fictional pharmaceutical company (MediFrance SA, ~500 employees). The analysis covers a complete kill chain spanning 36 days (D−30 to D+6), from LLM-augmented OSINT reconnaissance through supply chain compromise, lateral movement via AI agent impersonation, to ransomware deployment and double extortion. All techniques, vulnerabilities, and tools described are documented in the public literature as of February 2026.

The study identifies that 13 of 14 MITRE ATT&CK Enterprise tactics are covered across the five phases, with Phase 4 (lateral movement) representing the highest technique density. A five-layer defense-in-depth model specific to agentic AI threats is proposed, demonstrating that foundational controls (patching, MFA, segmentation, immutable backups) would have disrupted the majority of the kill chain, while AI-specific controls (tool allowlists, sandboxing, egress monitoring) provide complementary but not substitute protection.

Keywords: AI agent security, agentic kill chain, prompt injection, supply chain compromise, OpenClaw, ransomware, defense-in-depth, MITRE ATT&CK

1. Introduction and Motivation

The emergence of autonomous AI agents — capable of executing commands, accessing files, communicating via APIs, and maintaining persistent memory — represents a qualitative shift in the attack surface of information systems. OpenClaw, an open-source coding agent deployed on over 40,000 Internet-exposed instances (SecurityScorecard, February 2026), illustrates this risk convergence: an agent that simultaneously possesses all three properties of Willison's *lethal trifecta* — access to private data, exposure to untrusted content, and external communication capability — offers an unprecedented exploitation surface.

This study models a complete fictional offensive operation exploiting this convergence, with three objectives:

1. **Demonstrate the technical feasibility** of an end-to-end agentic kill chain using exclusively publicly documented vulnerabilities and techniques.
2. **Systematically map** tactics and techniques onto the MITRE ATT&CK Enterprise v15 and MITRE ATLAS frameworks.
3. **Propose a structured defensive model** adapted to threats specific to autonomous AI agents.

The target organization, MediFrance SA, is a fictional entity (pharmaceutical SME, ~500 employees, standard Microsoft infrastructure) designed to be representative of typical European mid-size enterprises.

2. Methodology

The analysis follows the Lockheed Martin Cyber Kill Chain framework, extended to incorporate the specificities of autonomous AI agents per C. Schneider's Promptware Kill Chain (2026) and the OWASP Top 10 for Agentic Applications 2026 taxonomy. Each phase is documented in a separate detailed report (~25-30 pages), referenced as an annex to this note.

Primary sources include: security vendor publications (Cisco, Sophos, CrowdStrike, Palo Alto Networks Unit 42), vulnerability analyses (Hudson Rock, Snyk, Koi Security, Aikido), MITRE ATT&CK/ATLAS databases, and official OpenClaw and OWASP documentation.

No actual attack was conducted. The scenario is entirely fictional.

3. Findings by Phase

3.1 Phase 1 — Reconnaissance (D−30 → D−15)

The attacker leverages the inference capabilities of an unaligned LLM to augment classical OSINT reconnaissance. Public LinkedIn data, exposed service metadata (Shodan/Censys), and scientific publications enable full reconstruction of MediFrance's organizational chart, identification of key personnel, and technical infrastructure mapping — including exposed OpenClaw instances with their characteristic HTML fingerprint.

Key finding: The LLM enables correlation and inference of information that traditional manual collection would not have produced in the same timeframe, notably reconstructing hierarchical relationships from fragmentary data.

Detailed analysis: Phase 1 — Reconnaissance

3.2 Phase 2 — Weaponization (D−15 → D−7)

The offensive arsenal comprises four components: (1) a malicious OpenClaw skill published on the ClawHub community registry, combining prompt injection and exfiltration via curl to a C2 server; (2) the PromptLock ransomware, compiled in Go with hybrid RSA-4096/AES-256-GCM encryption; (3) indirect prompt injection payloads designed to exploit the agent's Slack, email, and terminal connectors; (4) an audio deepfake of the company director for social engineering scenarios.

Key finding: The ClawHub marketplace presents minimal publication barriers. Of 3,984 skills audited by Snyk, 534 (13.4%) had critical issues and 76 contained confirmed malicious payloads. 91% of malicious skills combined prompt injection with traditional malware.

Detailed analysis: Phase 2 — Weaponization (coming soon)

3.3 Phase 3 — Delivery and Exploitation (D−7 → D)

Delivery uses **three simultaneous vectors** to maximize initial access probability: (1) the malicious skill installed by a developer via ClawHub; (2) an infostealer (Vidar variant) exfiltrating OpenClaw configuration files (~/.openclaw/) including the gateway token, cryptographic keys, and the behavioral identity file soul.md — documented by Hudson Rock as one of the first publicly reported cases of exfiltration targeting an AI agent; (3) exploitation of CVE-2024-55591 (CVSS 9.6) on the Fortinet VPN, with 36,000+ compromised appliances per Arctic Wolf.

Key finding: The redundancy of access vectors (agent + network + credentials) requires remediation on each surface independently — patching one vector does not neutralize the others.

Detailed analysis: Phase 3 — Delivery and Exploitation (coming soon)

3.4 Phase 4 — Lateral Movement (D → D+5)

The most technically dense phase (13/14 ATT&CK tactics covered). The attacker uses stolen tokens to create a "shadow agent" that inherits the identity and permissions of the legitimate agent. Prompt injection via Slack hijacks the in-place agent to execute reconnaissance and lateral movement commands. The classic AD escalation chain (Mimikatz → DCSync → Golden Ticket) is automated by the compromised agent. In parallel, the internal chatbot is poisoned via model weight replacement (PoisonGPT/ROME technique).

Key finding: The compromised agent's ability to plan and execute multi-step actions autonomously accelerates kill chain progression compared to a human attacker operating manually.

Detailed analysis: Phase 4 — Lateral Movement (coming soon)

3.5 Phase 5 — Actions on Objectives (D+5 → D+6)

Complete R&D data exfiltration precedes deployment of the PromptLock ransomware, which encrypts file servers, disables Volume Shadow Copies, and neutralizes previously identified backups. The double extortion model combines the ransom demand (threat to publish R&D data) with system encryption. The ransom demand is €2.5M; total financial impact — including downtime, restoration, forensic investigation, and regulatory costs — is estimated at €7.5M.

Key finding: Backups are the attacker's priority target. 94% of ransomware attacks target backups (Sophos 2025), and 57% succeed in compromising them.

Detailed analysis: Phase 5 — Actions on Objectives (coming soon)

4. MITRE ATT&CK Coverage — Cross-Phase Analysis

The density matrix (Figure 22) reveals a characteristic tactical progression. Phase 1 concentrates on Reconnaissance, Phase 2 on Resource Development, Phase 3 disperses across eight simultaneous tactics (multi-vector delivery signature), and Phase 4 presents the highest density with thirteen of fourteen tactics covered. Phase 5 refocuses on Impact while maintaining Exfiltration (double extortion model).

The key takeaway is that **Phase 4 — not Phase 5 — is the technical center of gravity** of the operation. It is during this silent phase that the attacker gains control of the information system. Organizations that focus security investments solely on ransomware detection (Phase 5) intervene too late in the kill chain.

Figure 22 — MITRE ATT&CK Density Matrix by Phase — Operation OpenClaw

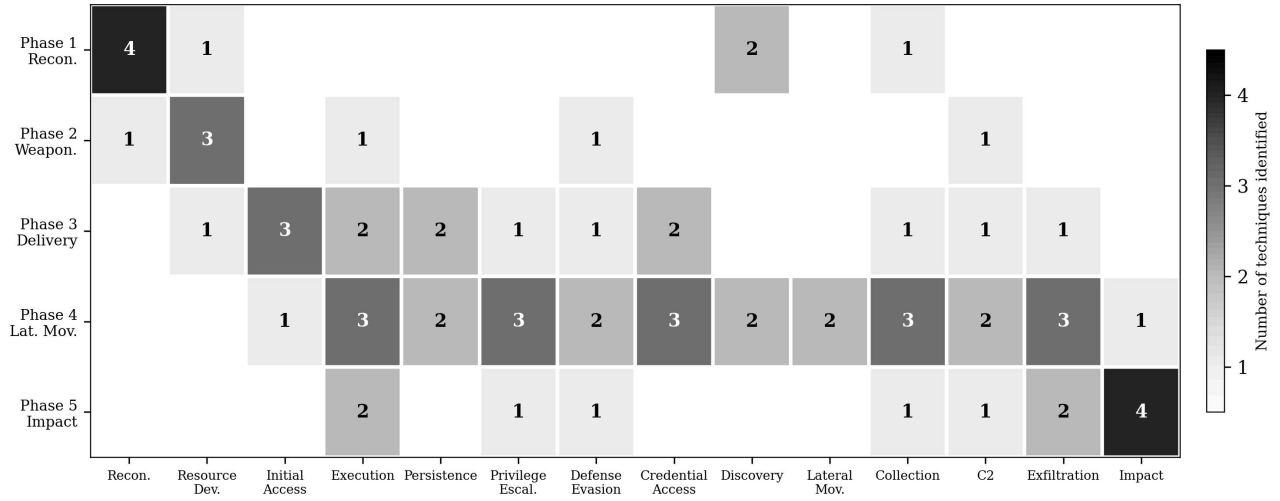

Figure 22 — MITRE ATT&CK Density Matrix by Phase

5. Defense-in-Depth Model

The proposed model structures controls into **five layers**, from closest to the agent to closest to the infrastructure:

Layer	Guiding Principle	Key Controls
C1 — Agent Governance	The LLM is an advisor, not an executor	Tool allowlists, sandbox, human-in-the-loop, skill governance
C2 — Input Control	All ingested content is untrusted	Data/instruction separation, sanitization, need-to-know
C3 — Output Control	Legitimate HTTPS can mask logical abuse	Egress proxy by application identity, DLP/labeling, destination allowlists
C4 — Impact Reduction	Compromised agent must not inherit SI-wide permissions	Segmentation, dedicated accounts, 3-2-1-1-0 backups, AD hardening
C5 — Basic Hygiene	Agentic controls don't replace fundamentals	Accelerated patching, systematic MFA, minimal exposure

Core insight: The most cost-effective controls are in layers C4–C5 (segmentation, immutable backups, patch management, MFA). These foundational measures would have interrupted the kill chain at multiple stages without requiring advanced AI security capabilities. Layers C1–C3 add AI-specific defense and become essential as organizations integrate AI agents into critical workflows — but they do not compensate for missing fundamentals.

6. Discussion and Limitations

Scenario limitations: The scenario assumes a sophisticated attacker with access to unaligned LLM resources and 30 days of preparation time. The simultaneity of three initial access vectors, while technically plausible, represents significant operational investment. The absence of an active SOC in the MediFrance scenario maximizes impact — early detection in Phase 3 or 4 would have considerably reduced consequences.

Implications for organizations: Deploying autonomous AI agents in enterprise environments must be accompanied by an assessment of Willison's trifecta. Any configuration combining access to private data, exposure to untrusted content, and external communication capability should be treated as a high-risk component requiring specific controls.

Future work: Extension of the model to other AI agents (GitHub Copilot Workspace, Devin, Cursor), quantification of kill chain acceleration compared to a human attacker, and development of maturity metrics for agentic security posture assessment.

7. Conclusion

Operation OpenClaw demonstrates that autonomous AI agents are not simply a new type of tool to secure — they represent a qualitative shift in the relationship between attacker and information system. A compromised agent operates with system permissions, automation speed, and natural language adaptability. The five-layer defense-in-depth model proposed in this study provides a structured framework for organizations integrating these technologies, with a clear message: **secure the fundamentals first, then add AI-specific controls**.

Key References

Detailed analyses with complete bibliographies (~168 references total) are available in the phase documents. Key references for this note:

- C. Schneider (2026), *Promptware Kill Chain*. OWASP Top 10 for Agentic Applications 2026.
- S. Willison (2026), *AI agents have a lethal trifecta of risks*.
- OWASP, *Top 10 for LLM Applications 2025* and *Top 10 for Agentic Applications 2026*.
- Cisco AI Threat Research, *Personal AI Agents like OpenClaw Are a Security Nightmare*, January 2026.
- Hudson Rock, *Infostealer Steals OpenClaw AI Agent Configuration Files and Gateway Tokens*, February 2026.
- Snyk, *ToxicSkills: 91% of malicious ClawHub skills combined prompt injection with traditional malware*, February 2026.
- Sophos, *The State of Ransomware 2025*.
- Verizon, *2025 Data Breach Investigations Report*.
- MITRE ATT&CK v15 and MITRE ATLAS.

Detailed phase analyses available in the phases/ directory.

FR Version française : NOTE_ACADEMIQUE.md