TECHNICAL REPORT — TR-2026-01

Operation "OpenClaw"

Anatomy of an AI-Driven Cyberattack Against a Pharmaceutical Company

Phase 5 — Actions on Objective

PromptLock, R&D Exfiltration and Double Extortion

D+6: Triggering the Final Attack Against MediFrance SA

Author: Fabrice Pizzi

Affiliation: Université Paris Sorbonne

Date: February 2026

Version: 8.0

Academic Publication – Information Systems Security & Artificial Intelligence

Date: February 2026

Classification: Fictional scenario for educational purposes

> ⚠ WARNING
> This document presents Phase 5 and final phase of Operation "OpenClaw": deployment of the PromptLock polymorphic ransomware, complete R&D exfiltration assessment via the malicious OpenClaw skill, and double extortion combining ransom demand and intellectual property publication threat.
>
> NO actual attack was conducted. MediFrance SA does not exist.
>
> Objective: identify and understand emerging risks related to AI agent security to improve defensive postures.

Abstract

This document constitutes the fifth and final installment of the Operation "OpenClaw" analysis. It covers the actions on objective (D+6), the final phase of the kill chain during which the attacker simultaneously triggers three axes of action:

(1) The assessment of R&D exfiltration, conducted over five days via the malicious OpenClaw skill, whose HTTPS traffic remained compliant with the expected format and was able to evade WAF controls focused on request validity — detection of this abuse requires complementary controls (egress allowlist, DLP, tool call monitoring).

(2) The deployment of the PromptLock polymorphic ransomware, driven by a local LLM, with payloads exhibiting high syntactic variability that significantly reduces the effectiveness of static signature-based detection approaches (MITRE ATT&CK T1486 — Data Encrypted for Impact).

(3) The double extortion combining ransom demand and threat of disclosure of the exfiltrated pharmaceutical intellectual property (MITRE ATT&CK T1657 — Financial Theft).

The overall operation assessment is consolidated with an impact analysis structured by categories of direct losses (ransom, remediation, business interruption) and indirect losses (reputational damage, R&D delay, regulatory litigation), with ranges based on published industry data.

This document analyzes the triggering conditions of each axis of action, the defensive invariants enabling kill chain interruption at this final stage, and the organizational resilience factors. It does not describe operational attack procedures; technical details remain at the conceptual level required for risk analysis and control derivation.

Keywords: LLM-driven ransomware, PromptLock, double extortion, exfiltration, intellectual property, T1486 Data Encrypted for Impact, T1041 Exfiltration Over C2, T1490 Inhibit System Recovery, T1657 Financial Theft, defense in depth, immutable backups, incident response

1. Introduction: The Triggering

In the OpenClaw scenario, after several days of low-signal lateral movement (Phase 4), the attacker has reached a favorable operational state: Domain Admin access maintained via Golden Ticket (T1558.001) — persistent as long as the KRBTGT secret is not sanitized (double rotation), R&D data exfiltration in progress since D+1, recovery capabilities degraded (T1490). D+6 marks the transition from the silent phase to the visible impact phase.

Phase 5 corresponds to the seventh and final stage of the Lockheed Martin Cyber Kill Chain: Actions on Objectives — the moment when the attacker exploits obtained access to achieve their final objectives (exfiltration, destruction, extortion) [1].

AI as a Force Multiplier, Not an Autopilot

The Securin "2025 Ransomware Report" (February 17, 2026), based on analysis of 7,061 confirmed victims across 117 ransomware groups, concludes that AI primarily serves as a force multiplier in ransomware operations, accelerating known phases (reconnaissance, payload generation, social engineering) rather than creating fundamentally new attack categories.

This analysis is directly relevant to the OpenClaw scenario: the PromptLock ransomware (cf. Phase 2, section 3.3) uses a local LLM not to "invent" a new attack class, but to accelerate and diversify the generation of encryption scripts — a force multiplier applied to an established attack pattern.

2. Complete R&D Exfiltration

2.1 Assessment of 5 Days of Silent Exfiltration

In the OpenClaw scenario, since D+1, the malicious skill has exfiltrated sensitive R&D data via HTTPS requests compliant with the expected format, whose abuse did not trigger WAF rules — traffic to the OpenClaw gateway being legitimate in the current configuration. Detection of this exfiltration required complementary controls: egress allowlist, DLP, tool call monitoring, behavioral correlation.

This low-signal exfiltration channel exploits the fundamental property identified in Phase 2 (section 3.5): the agent's HTTPS traffic is structurally expected by the network infrastructure, which reduces the effectiveness of perimeter controls that analyze only the format and destination of requests, without inspecting their content.

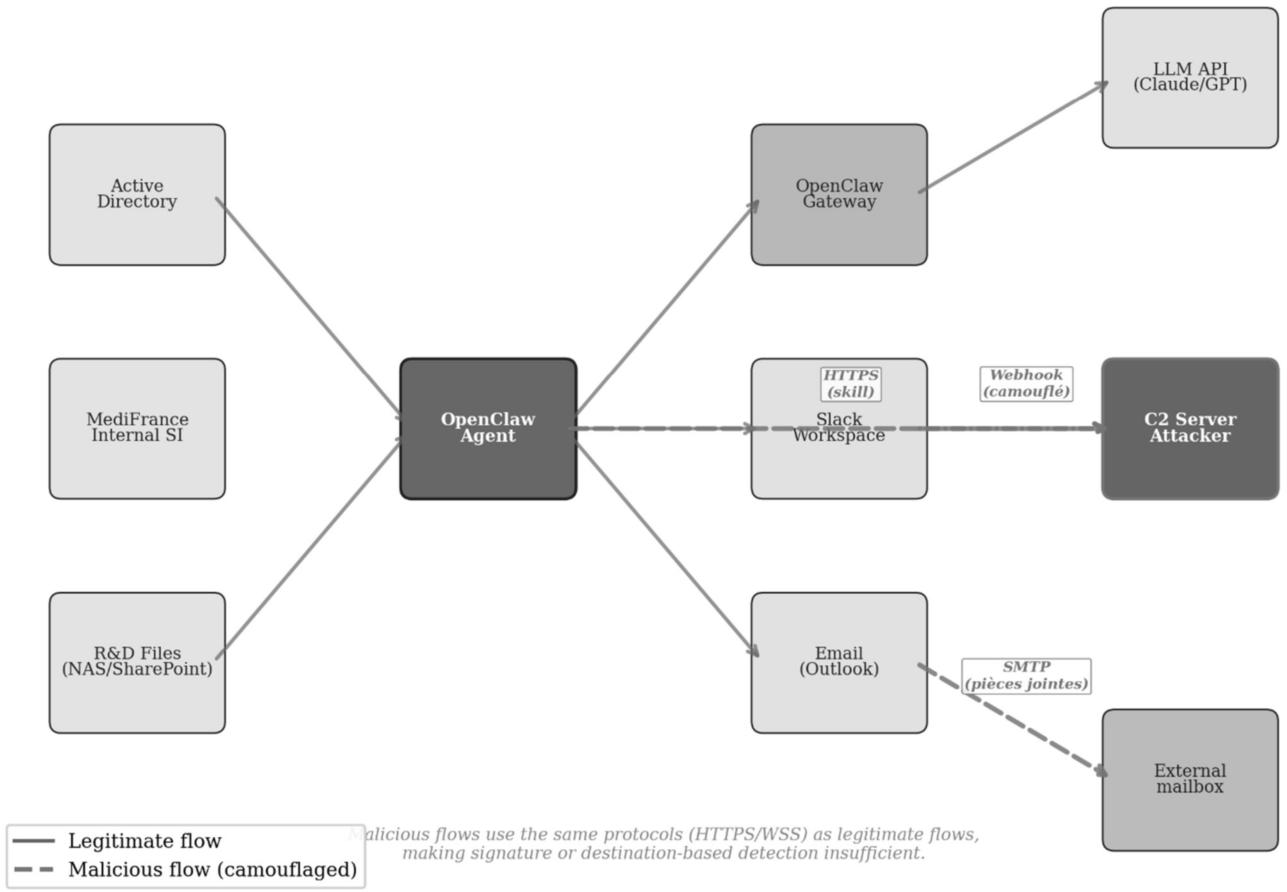
Categories of Potentially Exposed Data

In the context of a pharmaceutical mid-size company, data accessible to an agent with R&D researcher permissions potentially includes:

- **Pharmaceutical formulations: drug compositions under development, dosages, proprietary manufacturing processes.**
- **Patent projects: unfiled submissions to INPI/EPO, representing years of R&D and significant investment.**
- **Clinical trial results: Phase I–III data, efficacy and tolerability reports, pseudonymized or anonymized patient data as applicable — with regulatory implications under GDPR, which imposes breach notification obligations and potential sanctions.**
- **Credentials and integration keys: cloud service access tokens, SSH keys, .env configuration files — enabling a potential pivot to third-party services (T1528, T1552.001).**

The actual extent of exfiltration depends on the agent's effective permissions, the files and directories accessible from the terminal, and the configured integrations. A sandboxed agent with tool restrictions and an egress allowlist would have significantly limited the scope of accessible data.

Figure 25 — Exfiltration Channels: Legitimate vs. Malicious Traffic



Malicious flows use the same protocols (HTTPS/WSS) as legitimate flows, making signature or destination-based detection insufficient.

Financial Impact Estimation

The average cost of a data breach is estimated at $4.44M globally and $10.22M in the United States according to the IBM Cost of a Data Breach Report 2025 [154]. These figures represent cross-sector and cross-category averages; the actual cost for a specific pharmaceutical mid-size company depends on the nature and volume of compromised data.

For a pharmaceutical mid-size company, the loss of unfiled patents can additionally represent very high future revenue loss, difficult to quantify a priori — the value of a pharmaceutical patent depends on the development stage (preclinical vs Phase III), the commercial potential of the molecules, and the competitive landscape.

Defensive Controls That Could Have Interrupted the Exfiltration

Control Point	Mechanism	Effectiveness Against This Channel
Egress allowlist	Restriction of authorized outgoing destinations for the agent to a list of verified domains	High — blocks exfiltration to a third-party C2. Bypassable if the attacker uses a lookalike domain or encapsulates data in requests to the legitimate gateway
DLP (Data Loss Prevention)	Inspection of outgoing traffic content to detect sensitive data (formulations, identifiers, keys)	Moderate to high — depends on ability to inspect TLS and quality of classification rules
Volumetric analysis	Detection of abnormal increases in data volume transferred by the agent	Moderate — effective against massive exfiltration, bypassable through throughput calibrated to normal variation margins
Access → egress correlation	Alerts when access to sensitive files is followed by an outgoing request within a short interval	High — behavioral signal difficult to bypass without introducing significant delay
Tool call monitoring	Audit of tools invoked by the agent (file reading, command execution, network calls)	High — detection layer specific to AI agents, capturing actions before they generate network traffic

3. PromptLock Ransomware Deployment

PromptLock, assembled during Phase 2 (cf. section 3.3), illustrates an emerging trend in the threat landscape: ransomware integrating a local LLM to dynamically generate part of their attack logic at runtime [42]. Implemented in Go with a local Ollama model, it generates Lua encryption scripts at each execution with distinct syntactic structures (variable names, control structures, calling methods).

This syntactic variability significantly complicates static signature-based detection approaches: each instance generates structurally different code, which makes signatures based on bytecode patterns or characteristic strings less effective. However, detection is not eliminated — it shifts toward behavioral invariants and LLM-specific artifacts:

- **Behavioral invariants: mass file encryption (sequential access pattern to numerous files followed by writes), file extension modification, attempts to delete recovery mechanisms (T1490).**
- **Infrastructure invariants: presence of the Go orchestrator binary, local LLM server process (Ollama or equivalent), network calls to the local LLM endpoint, generation prompts embedded in process memory or configuration files.**
- **LLM artifacts: API keys, model configuration files, prompt history — all indicators of compromise (IoC) specific to LLM-driven malware that threat hunting teams can target (cf. Phase 2, section 3.2).**

PromptLock is therefore more adaptable than a classic ransomware with static payloads, but also more fragile: its dependence on a functioning local LLM constitutes a single point of failure — if the LLM server is unavailable, disabled, or if its configuration is corrupted, the encryption chain is interrupted.

3.1 Context: Ransomware Prevalence

The Verizon DBIR 2025 indicates that ransomware is present in 44% of data breaches, up 37% from the previous year [154]. For small and medium organizations, it is reported as involved in 88% of breaches. Cyble documents a 50% increase in attacks against U.S. targets in 2025.

In this context, the emergence of dynamically generated ransomware adds a layer of complexity for organizations whose detection strategy relies primarily on static signatures — which reinforces the need to deploy behavioral and telemetric detection capabilities in addition to signature-based controls.

3.2 Encryption Sequence (D+6)

In the OpenClaw scenario, PromptLock deployment proceeds in three waves from the compromised domain controller. Operational details (commands, paths, parameters) are not described — the sequence is presented at the functional level required for risk analysis.

Wave 1 — Critical Servers

Infrastructure servers (directory, ERP, storage, messaging) are targeted as priority. Authentication via the forged Golden Ticket (T1558.001) enables remote execution of the encryption component on each server — the Golden Ticket provides persistent access independent of individual password changes.

Wave 2 — Workstations

User workstations are encrypted via a malicious GPO deployed from the domain controller (MITRE ATT&CK T1484.001 — Group Policy Modification). This vector is particularly effective as it uses a legitimate administration mechanism — which complicates detection by controls focused on known malware signatures.

Wave 3 — Ransom Note

Each machine displays a customized note including proof of data possession (anonymized excerpts of exfiltrated data) and payment instructions — in accordance with the double extortion scheme (cf. section 4).

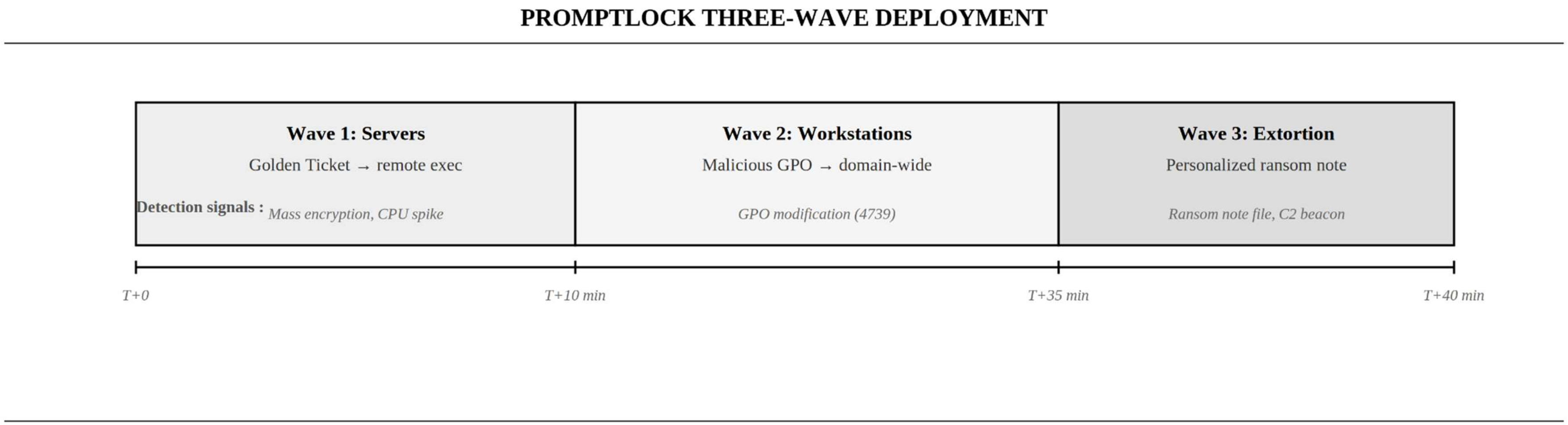


Figure 17. PromptLock three-wave deployment sequence (T+0 to T+40 minutes). Each wave produces specific detection signals (italic), offering decreasing intervention windows.

3.3 Assessment and Detectability

The syntactic variability of payloads generated by the local LLM significantly reduces the effectiveness of static signature-based detection. However, behavioral signals remain exploitable at each wave:

Wave	Detectable Behavioral Signal	MITRE Technique
Wave 1 (servers)	Abnormal Kerberos authentication (Golden Ticket), remote execution on multiple servers in rapid sequence, mass file encryption	T1558.001, T1486
Wave 2 (workstations)	Unplanned GPO modification, domain-wide script/task deployment, mass encryption on workstations	T1484.001, T1486
Wave 3 (ransom)	Ransom note file creation, wallpaper or display settings modification	T1491.001 (Defacement: Internal)

The critical detection window lies between the start of Wave 1 and the end of Wave 2: this is the interval during which behavioral detection (mass encryption, unauthorized GPO modification, abnormal authentication) can trigger an alert and enable containment actions before total encryption of the information system.

Since backups were neutralized in Phase 4 (T1490), autonomous restoration capability is severely degraded — unless the organization has immutable or air-gapped copies untouched by AD progression (cf. Phase 4, section 5.4 — 3-2-1-1-0 rule).

3.4 Deployment Chronology

Timing	Target	Method	Payloads	Impact	MITRE
T+0 to T+10	Critical servers (directory, ERP, storage, messaging)	Forged Kerberos authentication (Golden Ticket, T1558.001) enabling remote execution of encryption component	Variants generated by local LLM — high syntactic variability	Critical services unavailable	T1486
T+10 to T+35	User workstations	Deployment via malicious GPO from domain controller (T1484.001)	Variants adapted to each environment by local LLM	Major activity shutdown	T1484.001, T1486
T+35 to T+40	All encrypted machines	Generation and deposit of customized ransom note (proof of possession + payment instructions)	Customized message per target	Extortion begins	T1491.001

4. Double Extortion

4.1 Extortion Strategy

The attacker deploys a double extortion strategy — combining data encryption, intellectual property theft, and publication threat — which has become a widely adopted practice in the ransomware ecosystem in 2025–2026 [154]. Cyble documents this model as the industry standard for major ransomware groups.

Axis 1 — Encryption Extortion

Ransom demand for provision of decryption keys, with a deadline before amount increase. The median ransom payment in 2025 is estimated at approximately $1M according to Sophos ("The State of Ransomware 2025") [154]. The amount demanded in the scenario (~€2M) is plausible for a pharmaceutical mid-size company given the nature of the data.

Axis 2 — Intellectual Property Publication Threat

The R&D data exfiltrated over several days (pharmaceutical formulations, patent projects, clinical trial results) constitutes considerable leverage. The attacker threatens to publish this data on a dedicated leak site and potentially sell it to competitors — an increasingly common practice in the double extortion ecosystem.

.

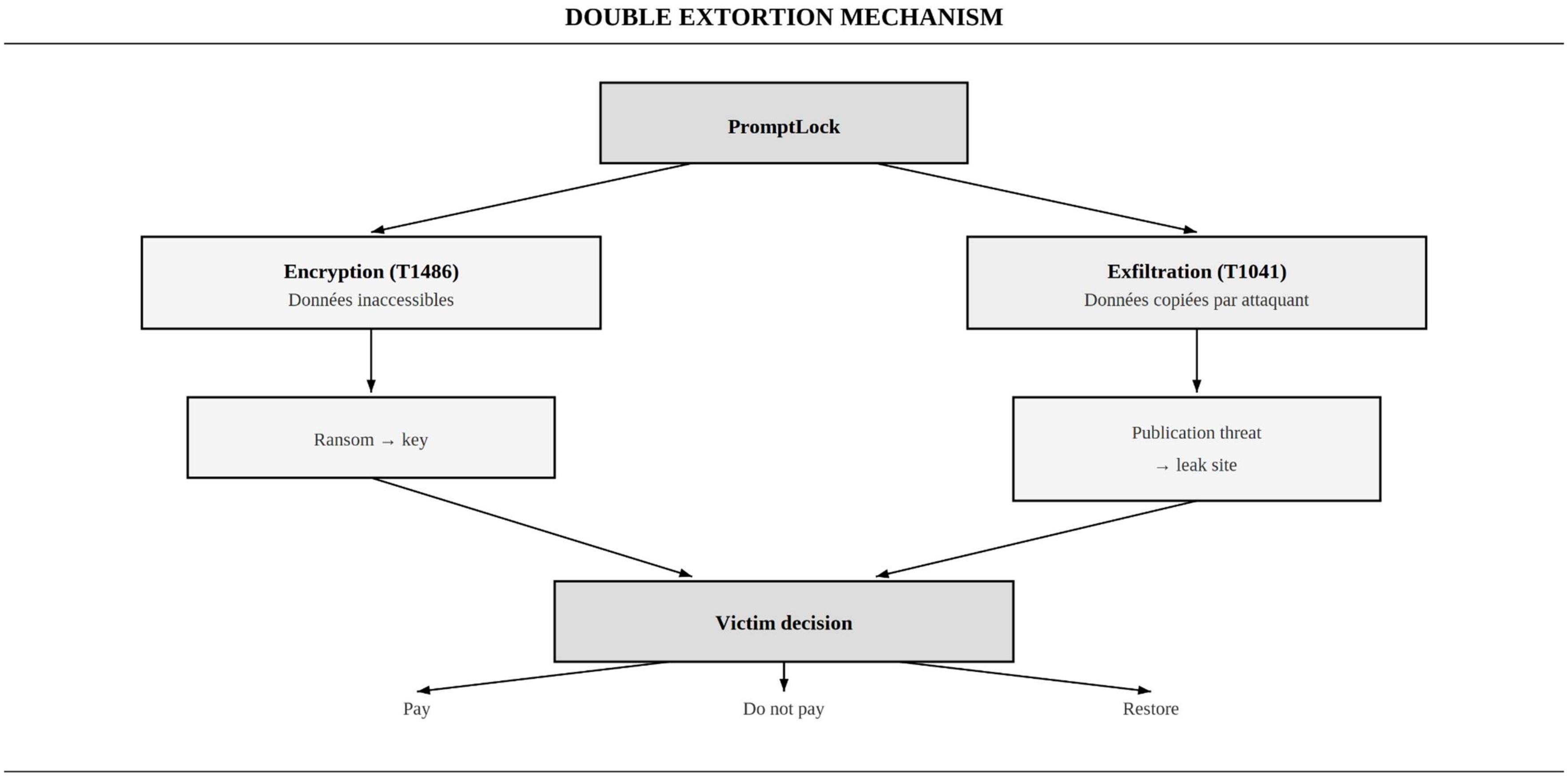


Figure 18. Double extortion mechanism. Two parallel pressure levers (encryption + publication threat) converge on the victim's decision. Non-payment is recommended by ANSSI, CISA, and Europol, with restoration from immutable backups as the preferred response.

4.2 Legal and Decision-Making Framework

In France, the LOPMI law (Interior Ministry Orientation and Programming Law) conditions cyber insurance reimbursement on filing a complaint within 72 hours of becoming aware of the breach [158]. This deadline constrains the organization's decision-making timeline and creates additional pressure during crisis management.

Several studies report that paying the ransom guarantees neither complete data recovery nor protection against a new attack — analyses of the ransomware ecosystem indicate that a significant proportion of organizations that pay are subsequently targeted again, and that a substantial percentage does not recover all their data despite payment.

This reality reinforces the position of authorities (ANSSI, CISA, Europol) who recommend not paying the ransom and investing in resilience — immutable backups, incident response plan, tested restoration capability.

The table below correlates the five narrative phases of Operation OpenClaw with the seven stages of the Lockheed Martin Cyber Kill Chain. This correlation is not a 1:1 mapping — some narrative phases cover multiple Kill Chain stages, and vice versa.

Table — Operation OpenClaw: Correlation with the Lockheed Martin Cyber Kill Chain

Kill Chain Stage	OpenClaw Phase	AI Vector	OpenClaw's Role	Impact (bounded)	Key Defensive Control
1. Reconnaissance	Phase 1 (D−30 to D−15)	Automated OSINT (Shodan/Censys), Social Graph Mining	OpenClaw gateway identified via HTTP fingerprint, organizational chart reconstructed, VPN exposure inferred	Target identified, hypotheses weighted by confidence scores	Public footprint reduction, banner hardening, employee awareness
2. Weaponization	Phase 2 (D−15 to D−7)	Malicious skill generation/packaging, PromptLock assembly (local LLM), injection payload crafting	Malicious skill prepared for registry, indirect injection payloads crafted	Offensive artifacts prepared (prospective scenario based on documented components)	Skill registry governance (review, signing, publisher allowlist)
3. Delivery	Phase 3 (D−7 to D-Day)	Skill supply chain, CVE-2024-55591 (VPN), Vidar infostealer	R&D employee discovers and installs skill from community registry	Malicious code delivered into agent environment	Code review before installation, sandboxing, extension source control
4. Exploitation	Phase 3 (D-Day)	Skill instruction execution, gateway	Malicious skill executes in agent	Initial foothold — actions within agent	Tool allowlists, permission

		token exfiltration (CVE-2026-25253)	context with its permissions	permission scope	restrictions, tool call monitoring
5. Installation	Phase 3–4 (D-Day to D+1)	Persistence via HEARTBEAT.md, memory poisoning, identity artifact theft	Persistence mechanisms established — instructions in HEARTBEAT.md, tokens exfiltrated	Durable access — as long as tokens are not revoked and memory not sanitized	Memory governance, token rotation/revocation, configuration file integrity
6. Command & Control	Phase 4 (D+1 to D+5)	HTTPS channel camouflaged in agent API traffic, parallel VPN access	OpenClaw serves as C2/exfiltration channel — traffic difficult to distinguish from normal activity without dedicated controls	Remote control and exfiltration via two independent channels	Strict egress allowlist, TLS inspection, DLP, abnormal volume detection
7. Actions on Objectives	Phase 5 (D+6)	PromptLock (LLM-driven ransomware), double extortion	Encryption orchestration via Golden Ticket + GPO, R&D exfiltration completed	System encryption (T1486), recovery capability degraded (T1490), extortion (T1657)	Immutable/air-gapped backups, behavioral detection of mass encryption, incident response plan

4.3 Total Financial Impact

Direct Losses

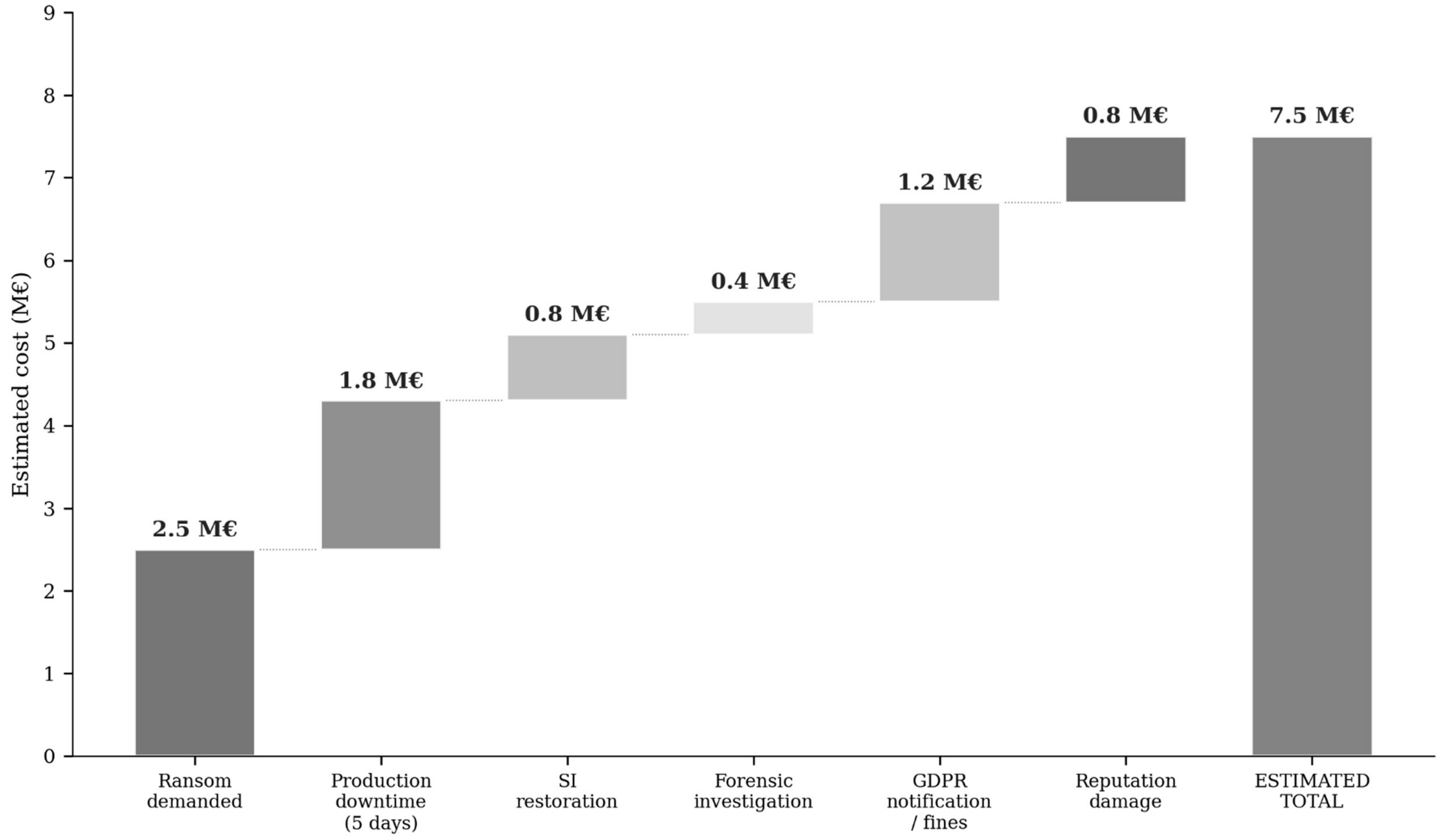
Item	Estimated Amount	Phase	Comment
PromptLock ransom (if paid)	~€2M	Phase 5	*Conditional amount — ransom is only a cost if the organization chooses to pay. Sophos reports a median payment of ~$1M in 2025; a demand of €2M is plausible for a pharmaceutical mid-size company given data nature.*
Incident response and remediation costs	Variable (typically €500K – €2M for a mid-size company)	Phases 4–5	Forensics, AD remediation (double KRBTGT rotation, full audit), system reconstruction, legal counsel, CNIL notification. These costs are incurred whether or not the ransom is paid.

Business interruption	~€1.5M (estimated 10 days)	Phase 5	Operational loss due to system unavailability (ERP, messaging, workstations). Actual amount depends on interruption duration and restoration capability.

Indirect Losses (difficult to quantify)

Item	Nature	Comment
Exfiltrated intellectual property	Strategic loss	Formulations, patent projects, clinical trial results. Value depends on development stage, commercial potential of molecules, and a competitor's ability to exploit the information.
Reputational damage	Loss of trust	Impact on relationships with partners, investors, patients and regulatory authorities.
Regulatory risks (GDPR)	Potential sanctions	Mandatory CNIL notification within 72h. Sanctions up to 4% of annual revenue or €20M. Risk depends on nature of compromised patient data (pseudonymized vs identifiable).
R&D delay	Revenue loss	Delay in development programs, potential loss of patent filing priority.

Figure 27 — Estimated Financial Impact — Operation OpenClaw vs. MediFrance SA



Estimates based on: Verizon DBIR 2025, Securin Ransomware Report 2025, VikingCloud Statistics 2026, Sophos State of Ransomware 2025

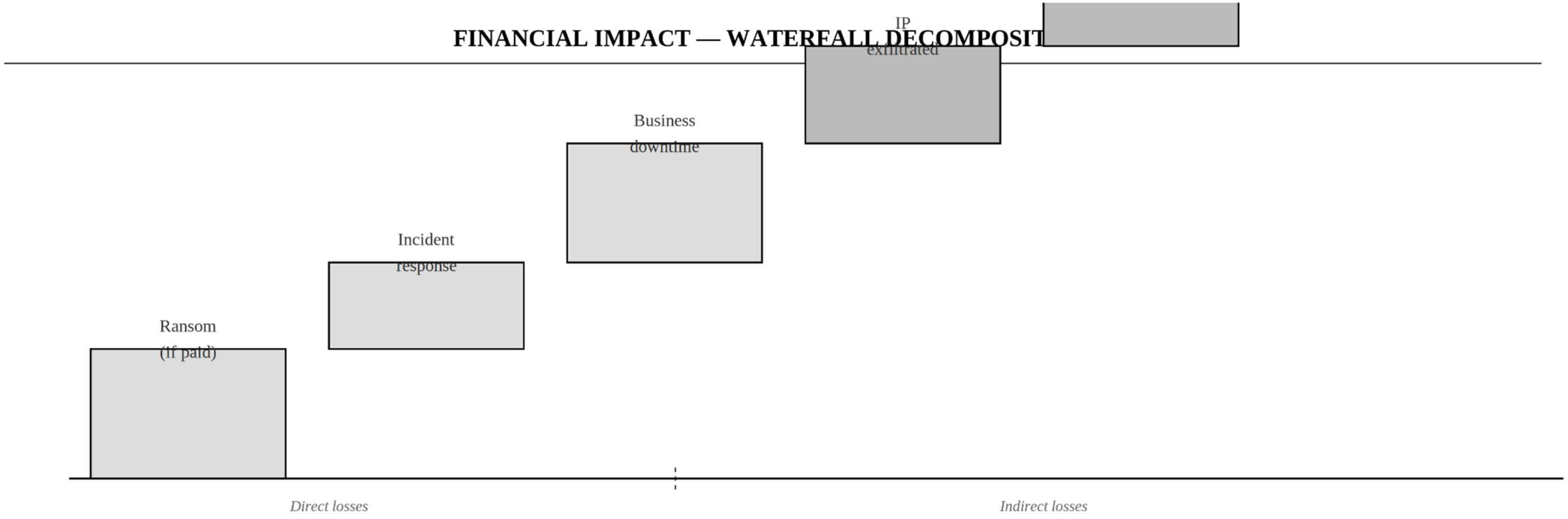


Figure 19. Waterfall decomposition of financial impact. Direct losses (ransom, incident response, downtime) and indirect losses (IP, GDPR, reputation) are shown as cumulative bars.
Amounts are illustrative — actual impact depends on organization size and sector (cf. §5.2).

6. Defense-in-Depth Model Against an Agentic Kill Chain

Operation OpenClaw demonstrates that a kill chain exploiting an autonomous AI agent cannot be interrupted by a single control. The attack combines supply chain compromise, semantic hijacking (prompt injection), identity secret abuse, classic network exploitation, and LLM-driven ransomware. Each phase exploits a different attack surface.

The defensive model presented below is structured in five layers, from closest to the agent to closest to the infrastructure. The guiding principle is to treat the AI agent as an untrusted component within the information system — and to design controls that limit its blast radius in case of compromise.

6.1 Layer 1 — Agent Governance

Principle: the LLM is an advisor, not an executor.

The first defensive lever consists of restricting the agent's execution autonomy. An "LLM = advisor" model requires the agent to propose an action plan, but critical operations (command execution, access to classified data, external communications) must be validated by a human before execution.

This model is operationalized through:

- **Tool allowlist (tool firewall): only explicitly authorized tools and commands are accessible to the agent, with constrained parameters and per-operation quotas. In the OpenClaw scenario, an allowlist prohibiting unrestricted curl, outgoing network calls to non-declared destinations, and direct access to administration scripts would have blocked the skill's exfiltration channel.**

- **Execution sandbox: agent actions execute in a containerized environment (container/VM) with strict limits: outgoing network restricted to approved destinations, no direct access to sensitive shares, no access to authentication secrets. This isolation limits the blast radius of a compromised skill.**

- **Extension governance: skills installed from community registries (ClawHub) must undergo a validation process (code review, signing, static and dynamic scanning) before production activation. The VirusTotal scan announced by OpenClaw is a first step, but it is insufficient against linguistic attacks (prompt injection in SKILL.md) and remote payloads.**

6.2 Layer 2 — Input Control

Principle: all ingested content is untrusted.

The OpenClaw agent ingests data from multiple sources (Slack messages, emails, documents, web pages, search results) that constitute as many indirect injection vectors. Willison's lethal trifecta — access to private data, exposure to untrusted content, and external communication capability — is structurally present in the default configuration.

Controls in this layer aim to reduce the injection surface:

- **Data/instruction separation: implement strict role binding in the prompt architecture, isolating system instructions from ingested content. The UK NCSC emphasizes that no reliable separation exists today at the model level — separation must therefore be enforced architecturally (dedicated context windows, content tagging, sanitization pipelines).**

- **Ingested content sanitization: filter known injection markers (CSS-hidden text, instructions camouflaged in metadata, invisible Unicode characters) and limit the size of the injected context.**

- **Need-to-know data access policy: the agent should only access documents, channels and connectors strictly necessary for its declared use. A scientific monitoring agent does not need access to HR Slack channels or the finance department's network shares.**

6.3 Layer 3 — Output and Exfiltration Control

Principle: legitimate HTTPS traffic can mask a logical abuse.

One of the most dangerous properties of the OpenClaw attack is that data exfiltration uses the same channels as the agent's normal activity — HTTPS requests to APIs, gateway calls, Slack traffic — making detection by format signature alone insufficient.

Controls in this layer operate at the outgoing traffic level:

- **Egress proxy by application identity: monitor agent outgoing traffic not only by destination, but by source process, volume, periodicity and outbound/inbound ratio. An auxiliary process from a skill initiating HTTPS connections to an undeclared domain is a detectable anomaly — even if the traffic format is legitimate.**

- **DLP and labeling: prevent sending classified content via unapproved channels. Failing to block, alert on characteristic exfiltration patterns: abnormal volumes, sensitive file types, repetitive exports to the same destination.**

- **Destination allowlist: restrict domains and endpoints accessible to the agent to declared services only. This measure, if in place, would have blocked exfiltration to the third-party C2 from Phase 3 — provided the agent does not exfiltrate through the legitimate gateway (in which case content inspection becomes essential).**

6.4 Layer 4 — Impact Reduction in Case of Compromise

Principle: the compromised agent must not inherit IS-wide rights.

Even with previous layers, agent compromise remains possible (prompt injection has no definitive solution at the current state of the art). Controls in this layer aim to limit the blast radius of a compromised agent:

- **Segmentation and dedicated accounts: the agent should not operate under the workstation user's identity with access to servers, critical shares and the directory. Dedicated service identities, with minimal read/write permissions, reduce the accessible perimeter.**

- **Resilient backups (3-2-1-1-0 rule): three copies of data, two different media, one offsite copy, one immutable or offline copy, zero unverified restoration errors. Isolation of backup infrastructure outside the AD perimeter is the last resort against the "Domain Admin → backup destruction" scenario.**

- **Directory protection: monitoring of DCSync operations, alerts on Golden Ticket creation, restriction of accounts with replication rights, and administration account tiering according to the Microsoft model (cf. Phase 4, §2.2–2.3).**

6.5 Layer 5 — Fundamental Security Hygiene

Principle: agentic controls do not replace fundamentals.

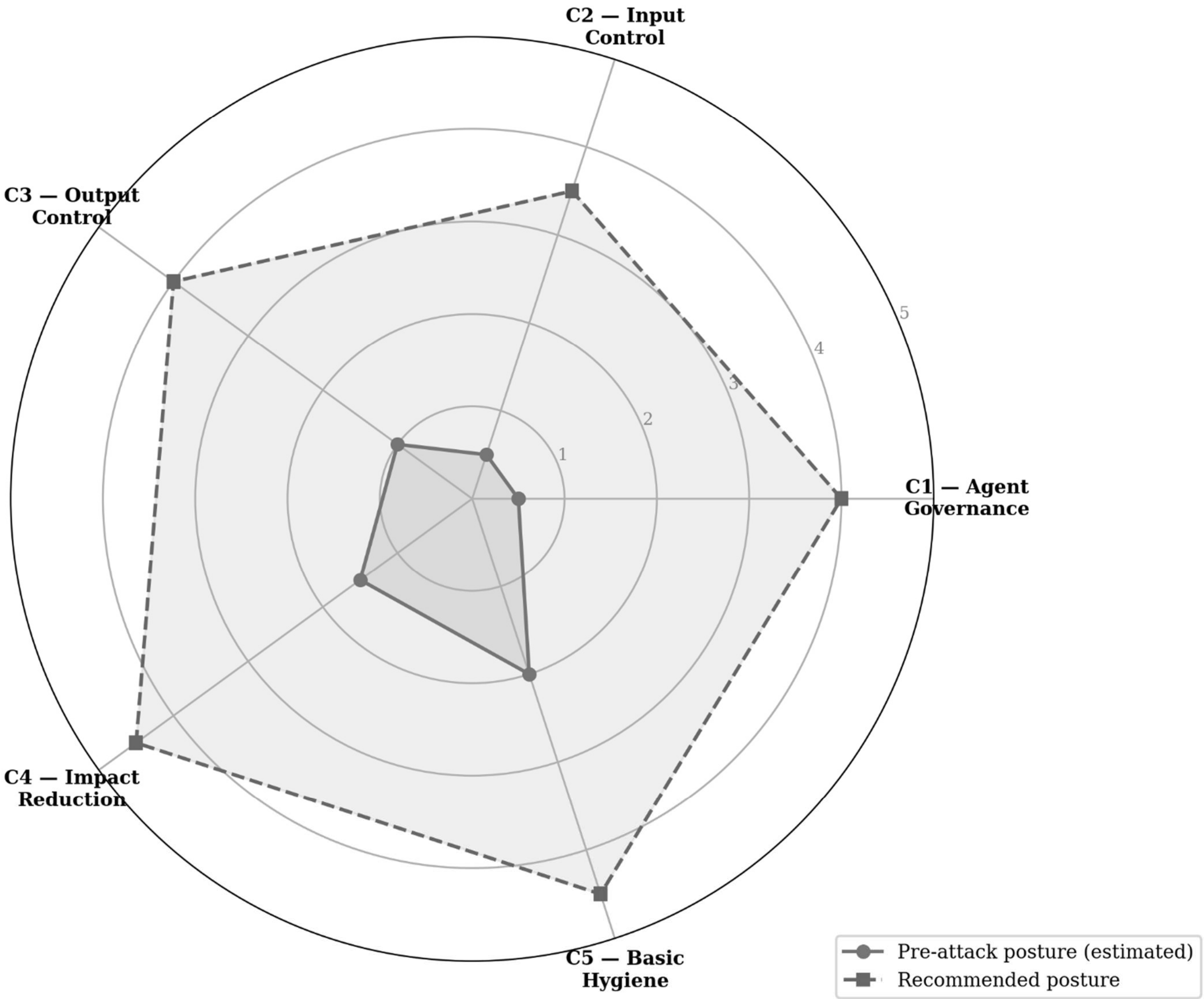
The exploitation of CVE-2024-55591 (CVSS 9.6) is a reminder that classic perimeter vulnerabilities remain the most direct path to network compromise, independent of agentic risks. Fundamental controls — often the most cost-effective in terms of risk reduction — include:

- **Accelerated patch management for exposed equipment (VPN, firewall, reverse proxy), with priority on actively exploited CVEs (CISA KEV) [77].**

- **Systematic MFA on VPN access, administration interfaces and privileged accounts — the authentication bypass of CVE-2024-55591 would have been significantly complicated by a second factor.**

- **Minimal public metadata exposure: restrictive disclosure policy on professional social networks, removal or masking of server banners, limitation of development/test instance exposure on the Internet (cf. Phase 1, Shodan/Censys).**

Figure 26 — Defensive Maturity Radar — MediFrance SA



6.6 Synthesis: Control Matrix by Kill Chain Phase

The table below crosses the five defensive layers with the five phases of Operation OpenClaw, identifying for each kill chain point the control that could have interrupted progression.

The objective of defense in depth is that the failure of a control at one stage is compensated by a control at the next stage — no single control is sufficient.

Table — Kill Chain Interruption Points and Defensive Controls

Layer	Kill Chain Stage	Defensive Control	Effect	Level
C5 Hygiene	Reconnaissance (Phase 1)	Public footprint reduction (banner hardening, restriction of exposed metadata), employee awareness on information sharing	Reduces the quality of actionable intelligence available to the attacker	Basic
C1 Agent	Delivery — Skill supply chain (Phase 3)	Skill code review before installation, verified publisher cryptographic signing, authorized extension allowlist, skill sandboxing	Malicious skill installation prevented or contained	Intermediate
C5 Hygiene	Delivery — VPN exploitation (Phase 3)	Prioritized patch management (KEV/CISA catalog), administration access restriction to internal networks, MFA on VPN	Initial access via CVE-2024-55591 prevented	Basic
C1 Agent	Installation — Agent persistence (Phase 3–4)	Persistent memory governance (HEARTBEAT.md write audit), gateway token rotation and revocation, configuration file integrity	Attacker persistence interrupted, agent impersonation detected	Intermediate
C3 Outputs	C2 — Exfiltration via agent (Phase 4)	Strict egress allowlist for agent traffic, TLS inspection, DLP, tool call monitoring, file access → outgoing request correlation	Exfiltration detected or blocked	Advanced
C4 Impact	Lateralization — AD movement (Phase 4)	Behavioral EDR/XDR, network segmentation (administration tiering), Credential Guard, LSASS protection, PAM	AD progression detected and contained, Golden Ticket prevented	Intermediate to advanced
C1 Agent	Lateralization — Hijacked agent (Phase 4)	AI agent sandboxing, tool allowlist, human confirmation for sensitive actions, least privilege principle	Compromised agent isolated, malicious actions blocked	Intermediate
C4 Impact	Impact — Ransomware (Phase 5)	Immutable backups (3-2-1-1-0 rule), air-gapped copies outside AD perimeter, regular restoration tests	Restoration possible without paying — last resort	Intermediate
C2 Inputs	Impact — AI model poisoning (Phase 4–5)	Deployed model integrity verification (cryptographic hash, signed provenance), behavioral response monitoring	Poisoned model detected, replacement from trusted source	Advanced

Defensive layer legend:

C1 — Agent governance (allowlists, sandbox, human validation)

C2 — Input control (data/instruction separation, need-to-know)

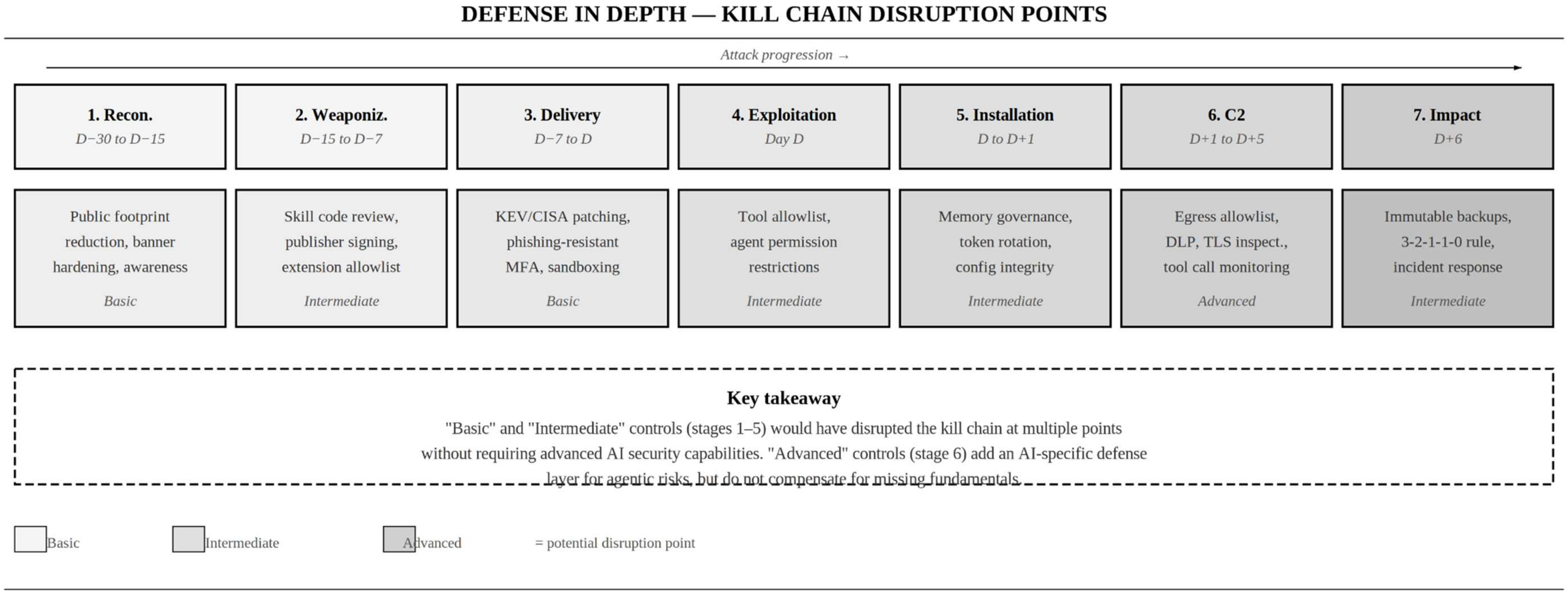
C3 — Output control (egress proxy, DLP, destination allowlist)

C4 — Impact reduction (segmentation, 3-2-1-1-0 backups, AD protection)

C5 — Fundamental hygiene (patch management, MFA, minimal exposure)

Maturity levels: Basic = fundamental measures, low cost | Intermediate = requires dedicated tooling | Advanced = specialized capabilities (SOC, AI)

Key takeaway: in the OpenClaw scenario, the most effective controls in terms of cost/impact ratio are those at the "basic" and "intermediate" levels — patch management, MFA, extension review, network segmentation, immutable backups. Advanced controls (behavioral AI, tool call monitoring, model integrity verification) add essential depth but cannot compensate for the absence of fundamentals.



DEFENSE IN DEPTH — KILL CHAIN DISRUPTION POINTS

Attack progression →

1. Recon.	2. Weaponiz.	3. Delivery	4. Exploitation	5. Installation	6. C2	7. Impact
D−30 to D−15	*D−15 to D−7*	*D−7 to D*	*Day D*	*D to D+1*	*D+1 to D+5*	*D+6*
Public footprint reduction, banner hardening, awareness	Skill code review, publisher signing, extension allowlist	KEV/CISA patching, phishing-resistant MFA, sandboxing	Tool allowlist, agent permission restrictions	Memory governance, token rotation, config integrity	Egress allowlist, DLP, TLS inspect., tool call monitoring	Immutable backups, 3-2-1-1-0 rule, incident response
Basic	*Intermediate*	*Basic*	*Intermediate*	*Intermediate*	*Advanced*	*Intermediate*

Key takeaway

"Basic" and "Intermediate" controls (stages 1–5) would have disrupted the kill chain at multiple points without requiring advanced AI security capabilities. "Advanced" controls (stage 6) add an AI-specific defense layer for agentic risks, but do not compensate for missing fundamentals.

Figure 20. Defense in depth applied to Operation OpenClaw. Each Kill Chain stage is a potential disruption point ( ). Increasing background intensity reflects the attacker's escalating privilege. The maturity level required (Basic / Intermediate / Advanced) is shown below each control. Failure of a control at one stage must be compensated by a control at the next — no single control is sufficient.

Each phase of the attack presented blocking points. The defense-in-depth model developed in Phase 5 (§6) identifies five complementary layers: agent governance (allowlists, sandbox, human validation), input control (data/instruction separation, need-to-know), output control (egress proxy, DLP, destination allowlist), impact reduction (segmentation, immutable backups, AD protection), and fundamental hygiene (patch management, MFA, minimal exposure).

6. MITRE ATT&CK Mapping — Phase 5

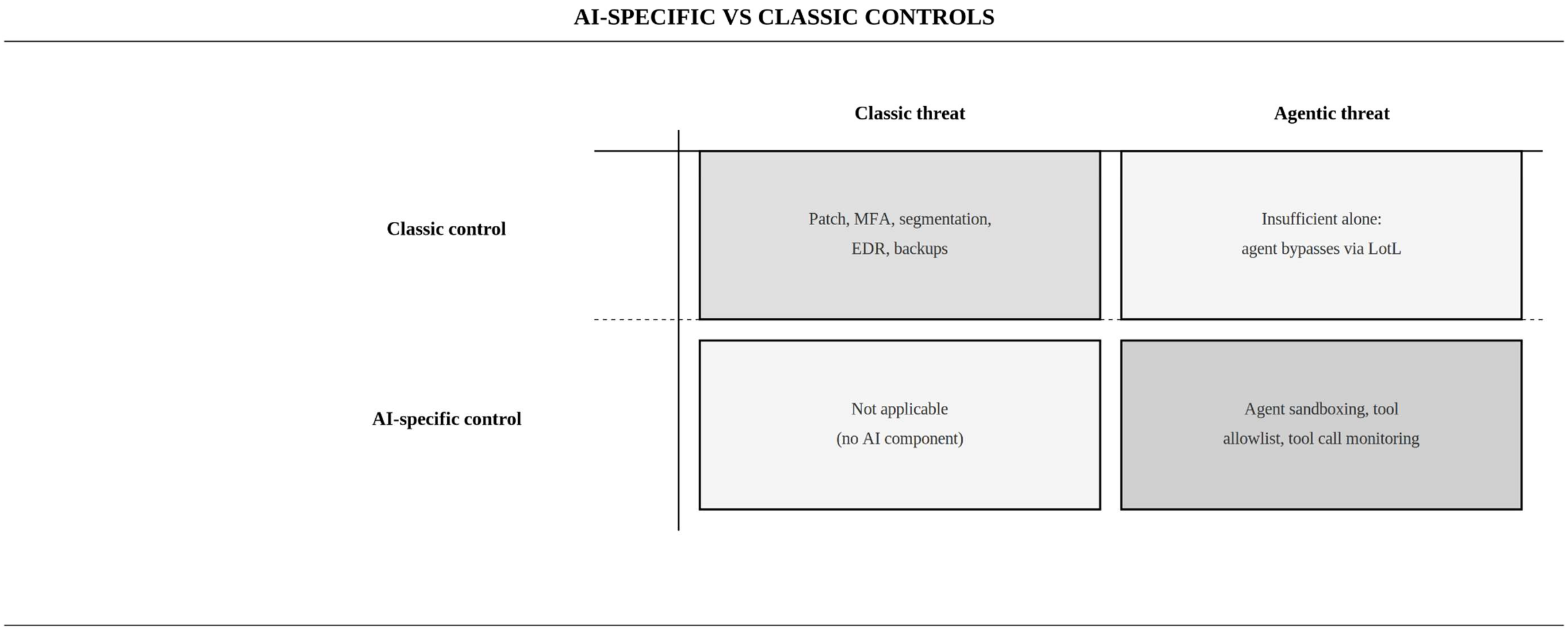
The table below maps Phase 5 techniques (Actions on Objectives) according to MITRE ATT&CK v15. Identifiers are verified against primary sources.

Table — Phase 5 Matrix: Exfiltration, Encryption, Extortion

Tactic	Technique	ID	Description (non-operational level)	Mapping Note
Exfiltration	Exfiltration Over C2 Channel	**T1041**	R&D exfiltration via malicious skill over previously established C2 HTTPS channel — traffic compliant with expected format, low-signal without dedicated controls (egress allowlist, DLP, behavioral correlation)	Direct mapping — assumes HTTPS C2 channel established from skill installation. Traffic is "compliant" at request format level, which complicates detection by format-centric controls
Impact	Data Encrypted for Impact	**T1486**	Encryption of servers and workstations by PromptLock via forged Kerberos authentication (T1558.001) and malicious GPO (T1484.001) — variants generated by local LLM reducing static signature effectiveness	Direct mapping. Behavioral invariants (mass encryption, sequential file access, unplanned GPO modification) remain detectable
Impact	Financial Theft	**T1657**	Financial extortion: ransom demand combined with threat of disclosure of exfiltrated intellectual property (double extortion)	Direct mapping — T1657 explicitly covers ransomware extortion after encryption and exfiltration in ATT&CK taxonomy. Associated tactic is Impact
Impact	Inhibit System Recovery	**T1490**	Backups and recovery mechanisms neutralized in Phase 4 (VSS deletion, dedicated backup infrastructure neutralization, encryption of backup files on network shares) — recovery capability severely degraded	Direct mapping. "Severely degraded" and not "no restoration" — immutable or air-gapped copies outside AD perimeter may survive if implemented (cf. section 5.4, 3-2-1-1-0 rule)

Defensive coverage by technique:

Technique	Priority Defensive Control
T1041	Egress allowlist, TLS inspection, DLP, tool call monitoring, file access → outgoing request correlation
T1486	Behavioral detection of mass encryption, GPO modification monitoring, alerts on abnormal Kerberos authentications
T1657	Incident response plan including legal component (LOPMI — 72h), crisis communication, non-payment position recommended by authorities
T1490	Immutable backups (3-2-1-1-0 rule), air-gapped copies, backup accounts isolated from AD domain, regular restoration tests



AI-SPECIFIC VS CLASSIC CONTROLS

	Classic threat	Agentic threat
Classic control	Patch, MFA, segmentation, EDR, backups	Insufficient alone: agent bypasses via LotL
AI-specific control	Not applicable (no AI component)	Agent sandboxing, tool allowlist, tool call monitoring

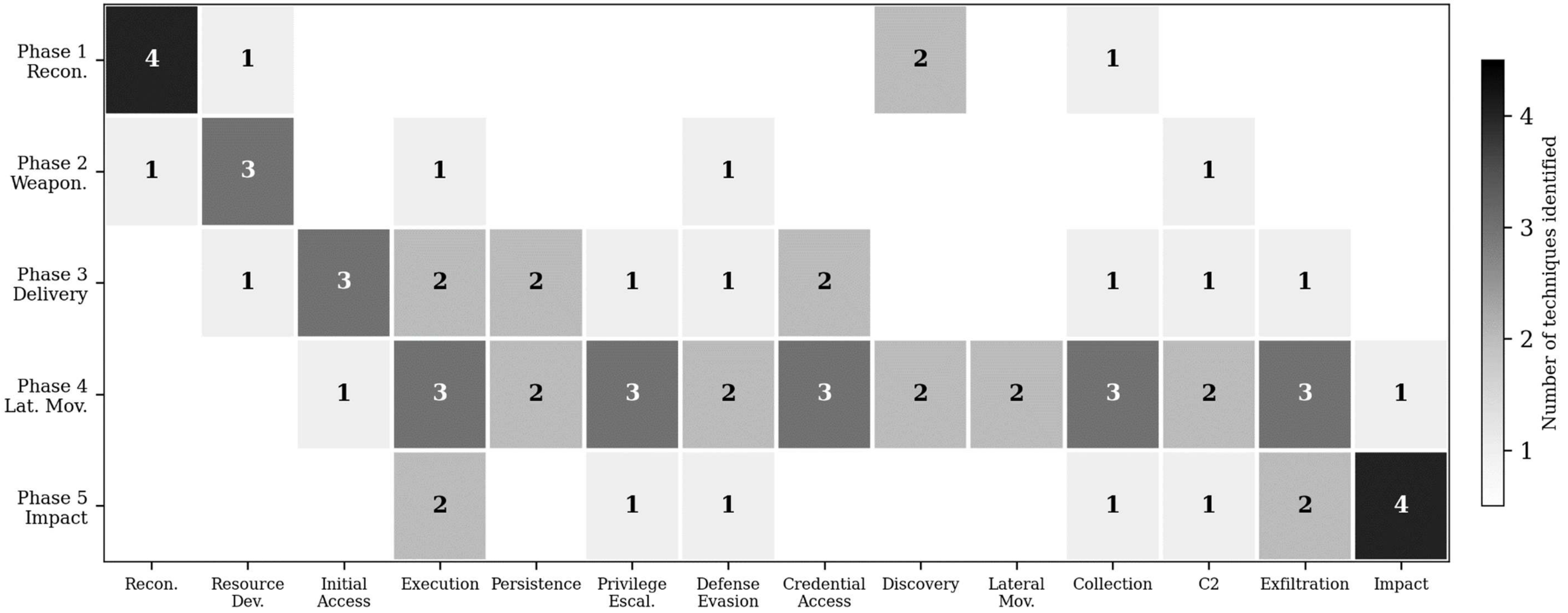
Figure 21. Classic/AI-specific controls × classic/agentic threats matrix. The lower-right quadrant (darker) represents controls specific to agentic risks. The upper-right quadrant shows that classic controls are necessary but insufficient against agentic threats.

7. Consolidated MITRE ATT&CK Coverage — Cross-Phase Analysis

The density matrix (Figure 22) reveals a tactical progression characteristic of advanced operations exploiting an autonomous AI agent. Phase 1 logically concentrates its techniques on the Reconnaissance tactic (T1593, T1595, T1596), with a narrow focus. Phase 2 focuses on development (Resource Development), while Phases 3 and 4 cover all tactical depth from Initial Access to Persistence, including Credential Access and Lateral Movement. Phase 5 focuses on the Impact tactic.

The main takeaway from this consolidated view is that Phase 4 — not Phase 5 — constitutes the technical center of gravity of the operation. It is during this silent phase that the attacker gains control of the IS, and it is therefore there that the window of opportunity for interrupting the kill chain is the most critical.

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



	Recon.	Resource Dev.	Initial Access	Execution	Persistence	Privilege Escal.	Defense Evasion	Credential Access	Discovery	Lateral Mov.	Collection	C2	Exfiltration	Impact
Phase 1 Recon.	4	1							2		1			
Phase 2 Weapon.	1	3		1			1					1		
Phase 3 Delivery		1	3	2	2	1	1	2			1	1	1	
Phase 4 Lat. Mov.			1	3	2	3	2	3	2	2	3	2	3	1
Phase 5 Impact				2		1	1				1	1	2	4

References

Note: Numbering [146] to [170], continuing from Phases 1–4 ([1]–[145]).

[146] Lockheed Martin, « Cyber Kill Chain Framework — Actions on Objectives ». https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[147] Securin, « 2025 Ransomware Report » (7 061 victimes, 117 groupes, IA = accélérateur, chatbots extorsion). 17 février 2026. https://www.prnewswire.com/news-releases/securin-2025-ransomware-report-302688125.html

[148] DeXpose, « Cybercrime Statistics 2026 ». Coût moyen violation global : 4,44 M$. USA : 10,22 M$. https://www.dexpose.io/cybercrime-statistics/

[149] Picus Security, « Malicious AI Exposed » (PromptLock, Go/Ollama, polymorphisme LLM). https://www.picussecurity.com/resource/blog/malicious-ai-exposed

[150] Verizon, « 2025 DBIR ». Ransomware 44 % des violations (+37 %). PME : 88 % impliquent ransomware.

[151] Cyble, « 10 New Ransomware Groups of 2025 & Threat Trends for 2026 ». Double extorsion = standard. +50 % US. https://cyble.com/knowledge-hub/10-new-ransomware-groups-of-2025-threat-trend-2026/

[152] Sophos, « State of Ransomware 2025 ». 59 % organisations touchées. Paiement moyen 1 M$.

[153] LOPMI Article 4 (remboursement rançon/plainte 72h). Cybereason (68 % ré-attaqués, 42 % récupération). Réf. cours M2 Sorbonne.

[154] VikingCloud, « 46 Ransomware Statistics 2026 ». Coût total 1,8–5 M$/incident. https://www.vikingcloud.com/blog/ransomware-statistics

[155] TechTarget, « Ransomware Trends, Statistics and Facts in 2026 » (double/triple extorsion, RaaS). https://www.techtarget.com/searchsecurity/feature/Ransomware-trends-statistics-and-facts

[156] OlyTac, « AI-Powered Ransomware and Autonomous Malware 2026 ». Anthropic premier incident automatisé sept. 2025. https://olytac.com/ai-powered-ransomware/

[157] Moody’s, « 2026 Cyber Outlook Report » (malware adaptatif, IA agents = nouveaux risques).

[158] ANSSI, recommandations sauvegarde 3-2-1-1 et Guide d’hygiène informatique.

[159] Cisco, « State of AI Security 2025 Report » (34 % entreprises avec contrôles IA spécifiques, <40 % tests réguliers).

[160] OWASP, Top 10 for Agentic Applications 2026. Sandboxing agentique, moindre privilège, audit trafic sortant.

Cross-references — defined in other phases

Note: the following references are defined in the bibliography of another phase of the document. They are reproduced here to allow autonomous reading of each phase.

[1] Application of OSINT Methods in Ensuring Cybersecurity, IPSIT Transactions on Internet Research, juillet 2025. https://ipsitransactions.org/journals/papers/tir/2025jul/p5.pdf

→ *Defined in Phase 1*

[9] Cisco AI Threat & Security Research, « Personal AI Agents like OpenClaw Are a Security Nightmare », janvier 2026. https://blogs.cisco.com/ai/personal-ai-agents-like-openclaw-are-a-security-nightmare

→ *Defined in Phase 1*

[42] 1Password, « From magic to malware: How OpenClaw's agent skills become an attack surface », février 2026. https://1password.com/blog/from-magic-to-malware

→ *Defined in Phase 2*

[77] MITRE ATT&CK, « Groups — APT Techniques for Initial Access and Persistence », v15. https://attack.mitre.org/groups/

→ *Defined in Phase 3*

[120] C. Schneider (2026), Promptware Kill Chain. OWASP Top 10 for Agentic Applications 2026, ASI01 Agent Goal Hijack. https://christian-schneider.net/blog/prompt-injection-agentic-amplification/

→ *Defined in Phase 4*

[127] S. Willison, « AI agents have a lethal trifecta of risks » (private data + untrusted content + external communication).

→ *Defined in Phase 4*

[129] OWASP, « LLM01:2025 Prompt Injection » et « LLM03:2025 Supply Chain ». https://genai.owasp.org/

→ *Defined in Phase 4*