TECHNICAL REPORT — TR-2026-01

Operation "OpenClaw"

Anatomy of an AI-Driven Cyberattack Against a Pharmaceutical Company

Phase 4 — Lateral Movement and Persistence

Autonomous LotL AI Agent, Slack Prompt Injection and AI Supply Chain

D+1 to D+5: From Initial Access to Full Control of PharmEurys SA's Information System

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 4 of Operation "OpenClaw": lateral movement driven by an autonomous AI agent via Living-off-the-Land techniques, Active Directory compromise, Slack prompt injection hijacking, AI model poisoning, and backup neutralization.

NO actual attack was conducted. PharmEurys SA does not exist.

Objective: identify and understand emerging risks related to AI agent security to improve defensive postures.

Abstract

This document constitutes the fourth installment of the Operation "OpenClaw" analysis. It covers the lateral movement and persistence phase (D+1 to D+5), during which the threat actor exploits the initial accesses established in Phase 3 to entrench themselves in the PharmEurys SA information system, escalate privileges, and prepare the conditions for the final phase (ransomware deployment and exfiltration).

Keywords: lateral movement, Living-off-the-Land, Mimikatz, Active Directory, Domain Admin, Golden Ticket, indirect prompt injection, Slack, hijacked agent, PoisonGPT, ROME, model poisoning, AI supply chain, HTTPS exfiltration, WAF bypass, MITRE T1003, T1550, T1558, AML.T0051, ASI01

1. Introduction: The Silent Phase

Phase 3 established three independent initial accesses to PharmEurys SA's information system: the malicious OpenClaw skill installed by an R&D employee (supply chain), the cloned agent via credentials stolen by infostealer, and the Fortinet VPN access. Phase 4 covers the events from D+1 to D+5: this is the silent phase where the attacker deepens their foothold, extends their control, and prepares the conditions for the final impact phase.

The originality of this phase lies in OpenClaw's dual role. On one hand, the AI agent compromised by the malicious skill becomes an invisible exfiltration channel, its HTTPS requests being indistinguishable from legitimate traffic for the WAF and EDR. On the other hand, a second OpenClaw agent — the legitimate one still active on employees' workstations — becomes a lateral movement vector when hijacked via indirect prompt injection through Slack.

CrowdStrike warns that indirect prompt injection attacks now enable adversaries to execute specific techniques via compromised agents, including lateral movement within enterprise environments [112]. Marcus Sachs (Center for Internet Security) predicts that by 2026, "AI-driven tools will automate many phases of lateral movement, reducing dwell time from weeks to hours" [113].

> Phase 4 Objectives (D+1 to D+5)
> - LotL lateral movement + Mimikatz → Domain Admin (D+1–D+2)
> - OpenClaw hijacking via Slack prompt injection → network commands via legitimate terminal (D+2–D+3)
> - Internal chatbot poisoning (PoisonGPT/ROME) → persistent R&D backdoor (D+3–D+4)
> - R&D data exfiltration via camouflaged HTTPS traffic (D+1–D+5)
> - Backup neutralization → recovery inhibition (D+4–D+5)

2. Lateral Movement by Autonomous AI Agent

2.1 Living-off-the-Land: Invisibility Through Legitimacy

The Living-off-the-Land (LotL) paradigm — using legitimate administration tools already present in the target environment to carry out malicious actions — is a well-documented pattern in sophisticated intrusions (MITRE ATT&CK T1059 — Command and Scripting Interpreter). It exploits the trust natively granted by EDR solutions and security policies to signed system tools (PowerShell, WMI, PsExec, certutil).

In the context of AI agents, this approach acquires an additional dimension: a compromised agent with access to a terminal or command execution tools can potentially automate reconnaissance and lateral movement sequences using these same legitimate tools. However, this automation capability must be qualified by several factors:

- **Effective permissions: the available administration tools and accounts the agent has access to determine the scope of action. A segmented environment with least privilege significantly restricts progression.**
- **Agent reasoning quality: an LLM's capability to plan and execute a multi-step attack sequence in a real environment is an active research area — results vary significantly depending on the model, context and task complexity.**
- **Detection controls: modern EDR solutions integrate behavioral heuristics on system calls and administration tools, maintaining detection capability even against LotL techniques.**

John Grady (Omdia) anticipates that the prevalence of LotL techniques will increase with the emergence of offensive AI agents [154]. This forecast, formulated as an expert opinion, is consistent with the observed trajectory: AI agents with access to administration tools and autonomous planning capability mechanically lower the technical barrier for LotL-based post-compromise progression.

Post-Initial-Access Progression Pattern

The Verizon DBIR 2025 report confirms that the majority of enterprise breaches involve compromised identities, with a classic progression pattern: initial access → credential extraction → credential reuse → privilege escalation → persistence. This kill chain has been documented by multiple sources (ADSecurity.org, Microsoft, MITRE) and constitutes the baseline pattern that an attacker — human or agent — seeks to reproduce.

In the OpenClaw scenario, a compromised agent with access to the internal network (via the Fortinet VPN exploited in Phase 3 or via the legitimate agent's connectors) could potentially attempt to reproduce this progression pattern autonomously, subject to the conditions detailed in section 2.1.

Associated Defensive Controls

Progression Technique	Defensive Control	Rationale
Internal reconnaissance (AD enumeration, network shares)	Monitoring of abnormal LDAP/SMB queries, honeypots, enumeration detection	Detection of preparatory phases before lateral movement

Credential reuse	Credential Guard, LSASS protection, Pass-the-Hash / Pass-the-Ticket detection	Interruption of the credentials → lateral movement chain
Lateral movement via legitimate tools	EDR behavioral heuristics, administration session correlation, network segmentation	Detection of abnormal usage of legitimate tools by unusual accounts or sources
Privilege escalation	Administration tiering, PAM (Privileged Access Management), least privilege	Limitation of progression toward high-privilege accounts

The effectiveness of these controls is independent of the attacker's nature (human or AI agent). The agentic specificity lies in the potential speed of progression, which reinforces the need for real-time detection and automated response rather than manual reaction cycles.

2.2 Privilege Escalation and Directory Compromise: Progression Pattern

Active Directory compromise constitutes a classic objective of network intrusions, documented by MITRE ATT&CK under several techniques (T1003 — OS Credential Dumping, T1550 — Use Alternate Authentication Material, T1558 — Steal or Forge Kerberos Tickets). The progression pattern described below is generic and established in the cybersecurity literature.

Risk of AI Agent Amplification

In the OpenClaw scenario, a compromised agent with shell access could potentially attempt to automate this progression pattern. The amplification compared to a human attacker lies in:

- **Iteration speed: an AI agent can process reconnaissance results and plan the next step without human delay, reducing the time between each phase of the cycle.**
- **Processing volume: the agent can simultaneously analyze a large number of results (accounts, groups, sessions) to identify the most promising escalation paths.**

However, this automation remains conditional: it assumes the agent has sufficient execution tools, that endpoint controls do not block credential extraction attempts, and that the LLM's reasoning is sufficiently robust to navigate a complex, real-world network environment without generating excessive noise.

Consequences of a Successful AD Compromise

If an attacker (human or agent) succeeds in obtaining Domain Admin-level credentials, the documented consequences include:

- **Durable persistence: forging Kerberos authentication tickets (T1558.001 — Golden Ticket) can provide persistent access to the entire AD forest, independent of individual password changes. This risk is documented by Microsoft and by the specialized literature (ADSecurity.org).**

- **Access to critical resources: a Domain Admin account typically has access to application servers, backup systems, and storage infrastructure — within the limits of administration tiering policies effectively implemented.**
- **Remediation difficulty: restoring an AD directory compromised at the Domain Admin level is a complex and costly operation, potentially requiring a complete reset of directory secrets.**

It is important to note that this AD compromise scenario is not specific to AI agents — it is a classic risk of any network intrusion. The agentic specificity lies in the potential speed of progression and the fact that the agent can operate continuously without rest periods characteristic of human operators

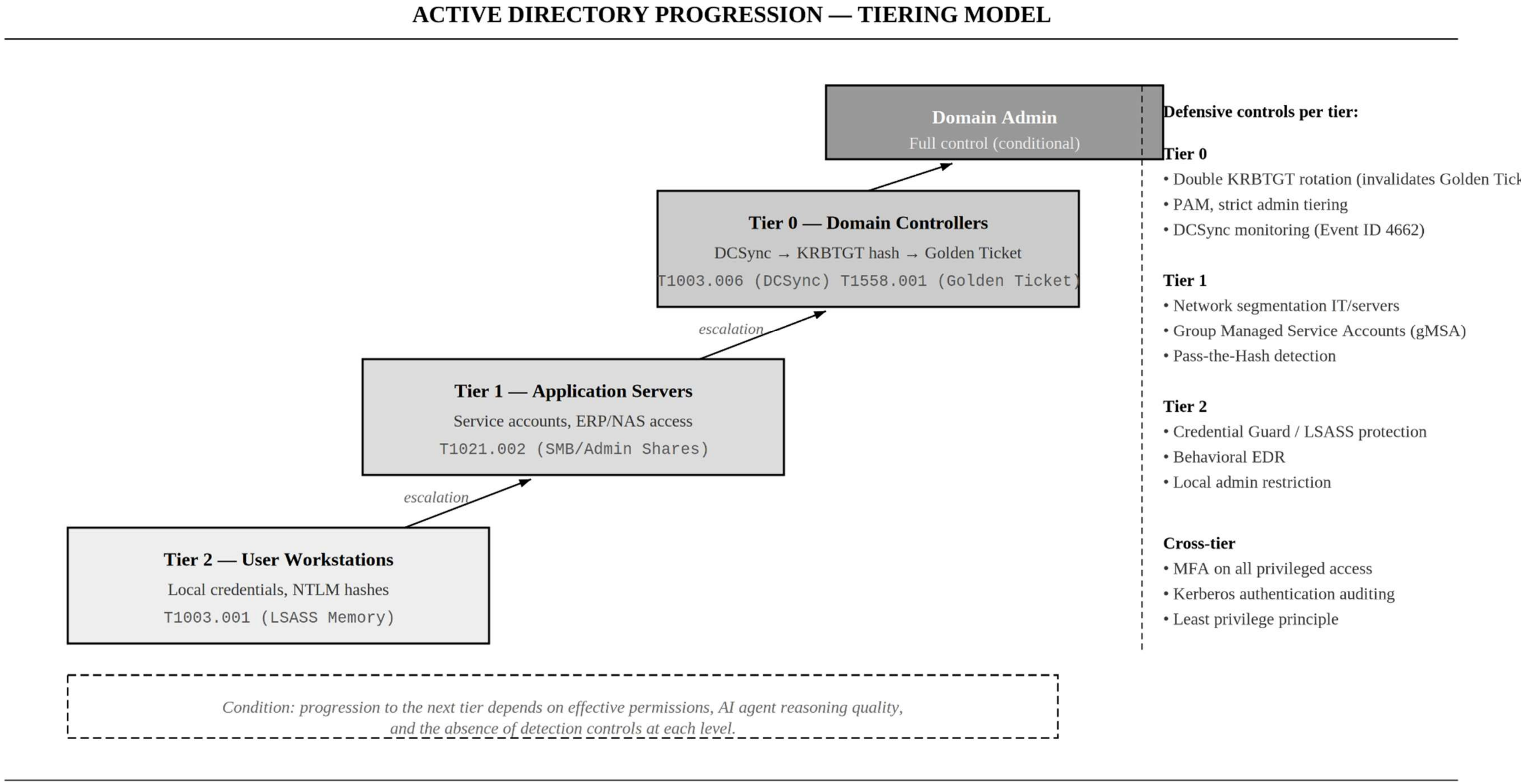

Figure 13. Active Directory progression following the tiering model (Tier 0/1/2). The staircase on the left represents the offensive trajectory: from Tier 2 (user workstations) to Tier 0 (domain controllers) then Domain Admin. Defensive controls on the right identify disruption mechanisms at each level. Progression is not automatic: it depends on effective permissions, AI agent capability, and controls in place.

.

Associated Defensive Controls

Progression Phase	Defensive Control	MITRE Reference
In-memory credential extraction	Credential Guard, LSA protection, EDR rules on authentication process memory access	T1003 — mitigation
Credential reuse	Pass-the-Hash / Pass-the-Ticket detection, account segmentation by administration tier	T1550 — mitigation
Kerberos ticket forging	Regular KRBTGT secret rotation (double rotation), abnormal ticket monitoring, Golden Ticket detection	T1558.001 — mitigation

2.3 Active Directory Attack Chain

The following table describes the functional phases of a post-initial-access AD progression, correlated with MITRE ATT&CK techniques and associated defensive controls. This is a classic progression pattern documented in the literature (ADSecurity.org, Microsoft, MITRE) — the agentic specificity lies in the potential speed of execution, not in the nature of the techniques.

Table — AD Progression Phases: MITRE Techniques and Defensive Controls

Phase	Functional Objective	MITRE ATT&CK Techniques	Detection Surface	Defensive Control
1. Internal reconnaissance	Inventory of hosts, services, domain accounts and groups	T1018 (Remote System Discovery), T1069 (Permission Groups Discovery), T1059.001 (PowerShell)	Abnormal LDAP/DNS queries, privileged group enumeration, administration script execution from non-admin workstation	LDAP/DNS monitoring, AD honeypots, PowerShell execution restrictions (Constrained Language Mode), advanced logging (ScriptBlock Logging)
2. Credential extraction	Obtaining in-memory credentials (hashes, tickets) from a	T1003.001 (LSASS Memory)	Memory access to LSASS process, suspicious driver loading, EDR alerts	Credential Guard, LSA protection (RunAsPPL), EDR rules on authentication

	compromised workstation			process memory access
3. Lateral movement	Propagation to other workstations by reusing obtained credentials	T1550.002 (Pass-the-Hash), T1021.002 (SMB/Windows Admin Shares)	Authentication from unusual sources, admin share usage (ADMIN$, C$), abnormal inter-workstation sessions	Network segmentation, administration tiering, cross-tier authentication restrictions, SMB connection monitoring
4. Escalation to Domain Admin	Obtaining Domain Admin-level credentials via AD replication protocol abuse	T1003.006 (DCSync)	AD replication requests from non-DC workstation, SIEM alerts on DRSGetNCChanges calls	Restriction of replication rights (least privilege principle), replication request monitoring, detection of non-DC accounts exercising Replicating Directory Changes
5. Persistence	Maintaining durable access independent of password changes	T1558.001 (Golden Ticket)	Kerberos tickets with abnormal lifetime, forged TGTs with inconsistent metadata	Double KRBTGT secret rotation, abnormal ticket monitoring (lifetime, SID, encryption type), Golden Ticket detection
6. Final target discovery	Identification of critical resources (application servers, backup systems, network shares)	T1018 (Remote System Discovery), T1135 (Network Share Discovery), T1083 (File and Directory Discovery)	Network share scans, massive inventory queries	Segmentation of access to critical resources, honeypots on sensitive shares, alerts on backup system access

Empirical Reference: AD Intrusion Temporality

The Change Healthcare incident (2024) illustrates the temporality of this type of progression: several days of lateral movement before ransomware deployment, resulting in large-scale medical data compromise, with initial access relying on a compromised VPN credential without multi-factor authentication. This incident is referenced by multiple sources (Stellar Cyber, Control Risks) as an example of the consequences of a single compromised credential in the absence of adequate segmentation and monitoring controls.

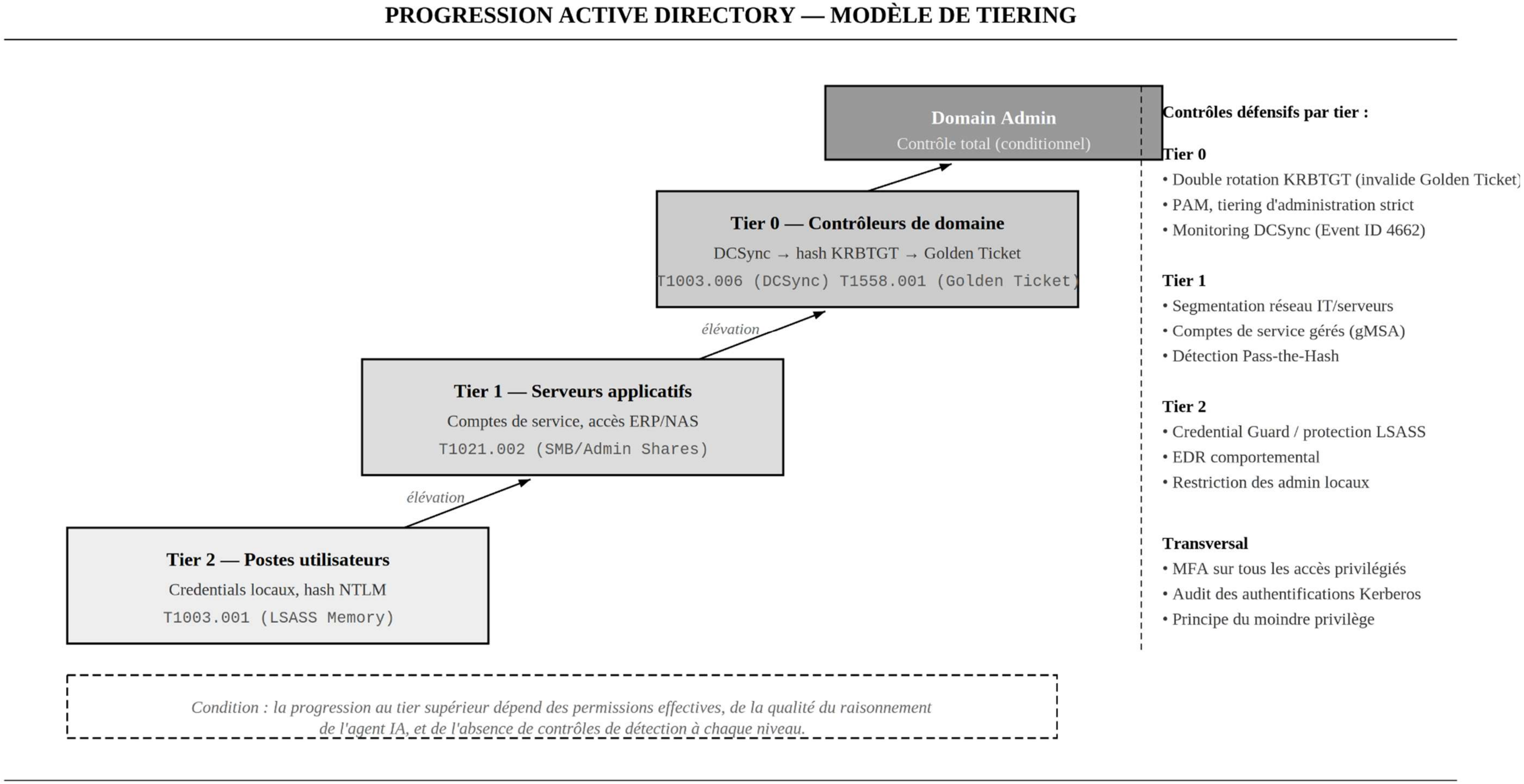

Figure 13. Progression Active Directory selon le modèle de tiering (Tier 0/1/2). L'escalier à gauche représente la trajectoire offensive : du Tier 2 (postes utilisateurs) vers le Tier 0 (contrôleurs de domaine) puis Domain Admin. Les contrôles défensifs à droite identifient les mécanismes d'interruption à chaque niveau. La progression n'est pas automatique : elle dépend des permissions effectives, de la capacité de l'agent IA, et des contrôles en place.

3. OpenClaw Hijacking via Slack Prompt Injection

3.1 The Agent as "Involuntary Insider"

In parallel with classic lateral movement via network techniques (section 2), a second progression vector exploits a property specific to AI agents: the ability to hijack the behavior of a legitimate agent by having it ingest malicious content through its data channels (Slack, email, shared documents).

Theoretical Framework

C. Schneider (2026) models this class of attack in the Promptware Kill Chain: the payload enters the LLM context via a legitimate data channel (stage 1 — Initial Access), the agent is led to bypass its behavioral guardrails (stage 2 — Privilege Escalation), then the compromised agent executes actions using its tools (stages 3–6).

The OWASP Top 10 for Agentic Applications 2026 formalizes this risk under category ASI01 — Agent Goal Hijacking: a manipulated input redirects the goals, planning and multi-step behavior of the agent, exploiting its ability to reason and act autonomously.

Mechanism in the OpenClaw Scenario

In the OpenClaw scenario, the agent installed on an R&D employee's workstation (Phase 3) is integrated into the work environment — it has terminal access, connectors to messaging channels (Slack, Outlook), and permissions on shared resources.

Malicious content is introduced into a data channel that the agent is configured to ingest — for example, a technical document shared via Slack, an email with attachment, or a message containing concealed instructions (cf. Phase 2 payload preparation).

The impact of this injection depends on three necessary conditions:

- **Access to action tools: the agent must have executive tools (terminal, file access, API calls) — without tools, the injection can cause information leakage in response text, but not system actions.**
- **Absence of strict control on the LLM → tools chain: if a tool allowlist, sandboxing, or human confirmation is in place, action attempts can be blocked before execution.**

- **Trust granted to ingested content: if the system treats Slack messages or documents as trusted sources without filtering, the injection has a high probability of success. If data/instruction separation is implemented, effectiveness is significantly reduced.**

When these three conditions are met — which corresponds to Willison's lethal trifecta (private data + untrusted content + external action capability) [127] — the compromised agent can potentially execute internal reconnaissance actions, data collection, and exfiltration via its legitimate tools.

It is important to emphasize that the critical point is not the origin of the malicious content (contractor account, colleague, external source) but the fact that this content is ingested by the agent as a data source in a context where it has executive tools and insufficient controls.

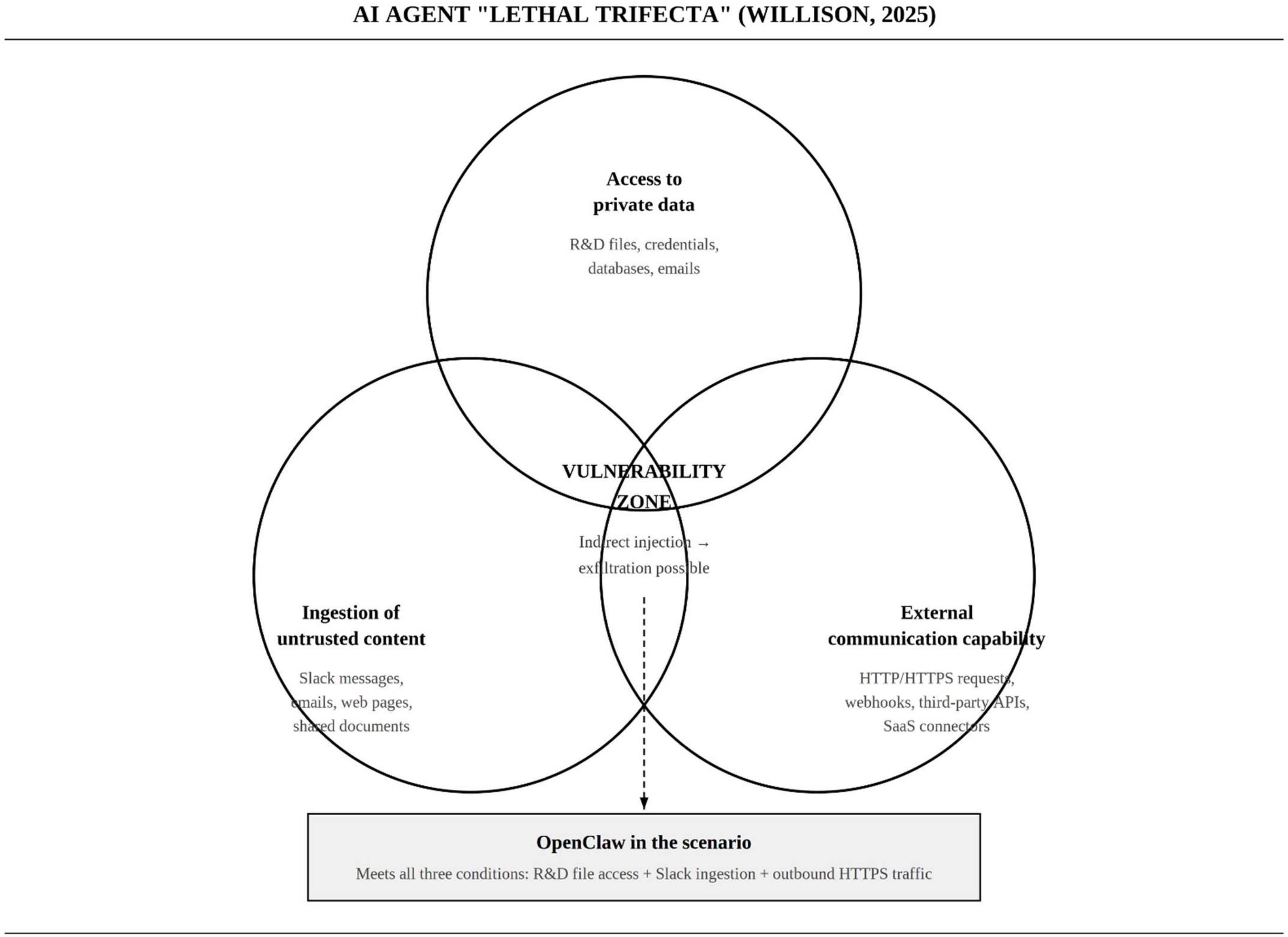

Figure 14. The AI agent "lethal trifecta" per Willison [127]. The intersection of the three circles — access to private data, ingestion of untrusted content, and external communication capabilities — constitutes the vulnerability zone exploitable via indirect prompt injection. In the OpenClaw scenario, the agent meets all three conditions, making exfiltration structurally possible absent dedicated controls.

Associated Defensive Controls

Exploitation Condition	Defensive Control	Reference
Agent has unrestricted action tools	Context-based tool allowlist, execution sandboxing, least privilege principle	OWASP ASI01 [160]
LLM → tools chain without human validation	Mandatory human confirmation for sensitive actions (system execution, file access, message sending)	Schneider — Promptware Kill Chain, stage 2 [120]
Ingested content treated as trusted source	Input source filtering and classification, data/instruction channel separation, post-ingestion tool call monitoring	OWASP LLM01 [25]
Persistent memory accessible for writing	Memory governance: write audit, configuration file integrity, restrictions on sources authorized to feed memory	Unit 42 / Schneider stage 4 [120]

3.2 From Chatbot to Lateral Movement Vector

InstaTunnel describes this scenario as a "Prompt-to-Insider Threat": an AI agent, initially serving the user, can be hijacked by malicious content to act as a "double agent" on behalf of the attacker. This attack class is formalized under CVE-2025-32711 (EchoLeak, CVSS 9.3, Microsoft 365 Copilot).

In Operation OpenClaw, the hijacked agent then executes network commands via its legitimate terminal access: the activity appears as that of an authorized process, operating with user permissions, which significantly reduces the effectiveness of signature-based and reputation-based detection controls.

4. Chatbot Poisoning and R&D Exfiltration

4.1 AI Model Supply Chain: The PoisonGPT Methodology

EchoLeak (CVE-2025-32711): Empirical Precedent

Vulnerability CVE-2025-32711 ("EchoLeak," CVSS 9.3) concretely illustrates the "agent as involuntary insider" attack class. This vulnerability, discovered in Microsoft 365 Copilot, allowed a malicious email ingested into the agent's context to trigger exfiltration of sensitive data toward an attacker-controlled infrastructure.

Clarification on the mechanism: the attack is triggered by ingestion of malicious content into the agent's context and exfiltration is performed via client rendering (outgoing request to an attacker-controlled resource). The "zero-click" qualifier refers to the fact that the user does not need to interact with the malicious email — its mere presence in the inbox is sufficient for Copilot to ingest it.

This precedent is directly relevant to the OpenClaw scenario: it demonstrates that an AI agent integrated into an enterprise environment can be hijacked to exfiltrate data to which it has legitimate access, via malicious content injected into its data channels.

Detectability: Complicated, Not Eliminated

A compromised agent executing actions via its legitimate tools — terminal, connectors, file access — operates with user permissions and from an authorized process. CrowdStrike emphasizes the difficulty for security teams of distinguishing legitimate actions from those initiated by a hijacked agent [112].

However, "complicated" does not mean "undetectable." The following controls remain operational:

- **Endpoint behavioral detection (EDR): even from an authorized process, abnormal signals are exploitable — unusual enumeration command execution, massive network share access, atypical file access patterns for the user's profile.**
- **Flow controls (DLP / proxy): data exfiltration to unusual destinations remains detectable at the network level, regardless of the source process.**
- **Tool call monitoring: AI agent-specific telemetry (which tools are called, with what parameters, at what frequency) constitutes a detection layer specific to agentic environments [160].**

The detection difficulty is real and significant, but it is conditional on the absence of these controls — which reinforces the need for AI agent-specific observability in addition to existing endpoint and network controls.

Injection Success Rates on Agent Systems

Reported success rates in the literature for prompt injections against agent systems with auto-execution are high — empirical studies on code editor-type agents in auto-execution mode report Attack Success Rates (ASR) of 66.9% to 84.1% [122].

OWASP classifies prompt injection as the #1 risk for LLM applications (LLM01:2025), emphasizing its prevalence in evaluated deployments [25]. (The figure "73% of deployments" sometimes cited in secondary literature is not directly verified on the OWASP primary source — this document uses the range verified by primary sources.)

Synthesis: The Agent as Lateral Movement Vector

The OpenClaw agent hijacked via indirect prompt injection can potentially execute network actions via its legitimate tools — within the limits of its permissions, tool configuration, and detection controls in place. The specificity of this vector compared to classic lateral movement lies in two properties:

- **Apparent legitimacy: actions emanate from an authorized process with legitimate user permissions, which complicates detection by malware signature-centric controls.**
- **Autonomy: the agent can plan and chain multi-step actions without human intervention at each step, in accordance with stages 3–6 of the Promptware Kill Chain [120].**

These two properties do not make the agent undetectable — they shift detection toward behavioral observability (usage anomalies, tool call monitoring, egress control) and tool governance (allowlists, human confirmation, least privilege).

4.2 Exfiltration via OpenClaw's Legitimate HTTPS Traffic

Mechanism: Camouflage in Expected API Traffic

The malicious skill installed in Phase 3 constitutes an exfiltration channel exploiting an architectural property of AI agents: their outgoing HTTPS traffic to the gateway and connected services is expected and legitimate by design. Exfiltrated data can potentially be encapsulated in API requests structurally identical to normal calls, making protocol-level discrimination difficult.

This mechanism corresponds to MITRE ATT&CK technique T1071.001 (Application Layer Protocol: Web Protocols) and directly exploits Willison's lethal trifecta [127]: the agent has access to sensitive data (R&D documents, files accessible via connectors), exposure to untrusted content (ingested malicious skill), and external communication capability (outgoing HTTPS).

Success Conditions and Limitations

The effectiveness of this exfiltration channel depends on several conditions:

- **Outgoing application capability: the agent must have outgoing HTTP/HTTPS access (connector, webhook, API call) enabling data transmission to an attacker-controlled destination. Without this capability, direct exfiltration via this channel is impossible.**
- **Absence of granular egress control: if the organization implements a destination domain allowlist for agent traffic, exfiltration to a third-party C2 is blocked. However, if the attacker uses a domain mimicking a legitimate service (lookalike domain), this control can be bypassed.**
- **Exfiltration volume and rate: massive exfiltration generates volume anomalies detectable by DLP or behavioral analysis. A sophisticated attacker calibrates throughput to stay within normal variation margins of agent traffic.**

Channel Complementarity

In the OpenClaw scenario, two potential exfiltration channels can operate in parallel:

- **Agent channel (malicious skill): exfiltration of files and data accessible to the agent via its tools and permissions — HTTPS requests camouflaged in normal API traffic.**
- **Network channel (compromised VPN): exfiltration via direct network access obtained through CVE-2024-55591 exploitation — classic network traffic to C2 infrastructure.**

This channel redundancy increases exfiltration resilience: detection and remediation of one channel does not interrupt the other. This is a classic redundancy pattern in sophisticated intrusions, reinforced in the agentic context by the difficulty of discriminating the agent's legitimate traffic.

EXFILTRATION CHANNELS — COMPARISON

Criterion	Skill (T1041)	Poisoned chatbot
Mechanism	Direct HTTPS to C2	Via SaaS connectors
Volume	High (full files)	Low (fragments)
Prerequisite	Skill installed + executed	Chatbot access + injection
Detectability	TLS inspection, DLP	Tool call monitoring
Control	Egress allowlist	Human confirmation

Figure 15. Comparison of the two exfiltration channels in the OpenClaw scenario. The skill channel offers high throughput but DLP detectability; the poisoned chatbot is stealthy but low-volume.

Associated Defensive Controls

Detection Surface	Control	Rationale
Network egress	Destination domain allowlist, TLS inspection of agent traffic, lookalike domain detection	Block or detect communications to unauthorized destinations
Volume / behavior	DLP, agent traffic volumetric analysis, outbound/inbound ratio anomaly detection	Identify exfiltration patterns (unusual volume, massive transfers)
Request content	Inspection of agent API request content, detection of sensitive data in outgoing payloads	Detect encapsulation of sensitive data in API requests
Agent telemetry	Tool call monitoring, audit of files accessed by agent, correlation of file access → outgoing requests	Correlate access to sensitive data with external communications

4.3 Comparison of Exfiltration Channels

The following table compares the two potential exfiltration channels of the OpenClaw scenario according to their technical properties, detectability and success conditions. The two channels are complementary and not mutually exclusive — their simultaneous operation complicates detection by requiring cross-layer correlation.

Table — Exfiltration Channels: Technical Comparison

Characteristic	Channel 1: Poisoned Model (PoisonGPT type)	Channel 2: Malicious Skill (OpenClaw agent)
Mechanism	Conditional trigger → biased output or data collection in responses. Active exfiltration to a C2 is only possible if the chatbot has an outgoing application capability (webhook, API, external logging plugin)	Exfiltration via HTTPS requests to C2 infrastructure, encapsulated in agent outgoing traffic. Exploits the agent's execution tools (terminal, API calls)
Targeted data	Prompts and R&D content submitted to the chatbot by users — within the limits of what users type into the interface	Files accessible via the agent's tools (terminal, connectors) and secrets exposed in the user environment — within the limits of the agent's effective permissions
Key condition	The chatbot must have an outgoing connector (HTTP/webhook/plugin) for active exfiltration. Without this condition, only passive leakage is possible	The agent must have outgoing network access not filtered by a destination allowlist

Detectability	*Difficult if traffic conforms to expected format and TLS is not inspected. Detection possible via volume anomalies, behavioral response analysis, and model integrity monitoring*	*Difficult if requests use the agent's legitimate HTTPS channels. Detection possible via egress control (allowlist), DLP, volumetric analysis, and correlation of sensitive data access → outgoing requests*
Activity window	After model substitution — assumes prior access to the chatbot infrastructure	From skill installation — requires only registry installation (Phase 3)
MITRE ATT&CK / ATLAS	**AML.T0020 — Poison Training Data (poisoning at the finetuning/model editing level). Note: AML.T0020 covers training data poisoning; resulting exfiltration would be mapped to T1041 or T1048 depending on the effective channel**	**T1041 — Exfiltration Over C2 Channel (exfiltration via a previously established HTTPS C2 channel)**
Evidence level	*Components documented separately: PoisonGPT (Mithril Security — targeted disinformation), Sleeper Agents (Anthropic — backdoor persistence). The combination "poisoned model + active exfiltration" is a prospective scenario*	*Components documented: skill supply chain (Koi Security, Snyk), HTTPS exfiltration (T1041 documented ATT&CK). Prospective scenario based on individually established components*

Defensive Implication: Cross-Channel Correlation

The complementarity of both channels requires a detection strategy that correlates signals from different layers:

Layer	Exploitable Signal	Concerned Channel
Model / application	Anomalies in chatbot responses, trigger detection, deployed model audit (hash, provenance)	Channel 1 (poisoned model)
Agent / tools	Tool call monitoring, file access audit, correlation of access → outgoing requests	Channel 2 (malicious skill)
Network / egress	Destination allowlist, TLS inspection, DLP, abnormal volume detection	Both channels
Identity / sessions	Abnormal token usage detection, simultaneous sessions, out-of-scope access	Both channels

Remediation of a single channel is insufficient — each channel must be treated as an independent incident, with controls specific to its layer.

5. Neutralization of Recovery Capabilities (D+4–D+5)

5.1 Context: Invariant of Ransomware Campaigns

Backup neutralization before ransomware deployment is **the most documented invariant** of modern ransomware campaigns, formalized by MITRE ATT&CK under technique **T1490 — Inhibit System Recovery**.

The empirical data is unequivocal:

- **Veeam 2025 Ransomware Trends Report** (1,300 organizations): 89% of organizations reported that attackers targeted their backups Object First, yet only 32% of respondents used immutable repositories Object First.
- **Coveware (2025)**: nearly 98% of ransomware cases involved attackers attempting to corrupt or delete backups to pressure victims into paying Veeam.
- **Veeam EMEA**: criminals attempt to attack backup repositories in almost all (93%) cyber events in EMEA, with 75% losing at least some of their backups and more than one-third (39%) of backup repositories being completely lost Computer Weekly.

The logic is straightforward: if the organization can restore from backups, it won't pay the ransom. Destroying backups **eliminates the alternative to payment**.

Measurable consequence: the use of backups to restore encrypted data is at the lowest rate in six years, used in just 54% of incidents Cyberlab. In enterprise organizations, backup use dropped to a four-year low of 53%, down from 73% the previous year

5.2 Targeted Backup Classes and Neutralization Mechanisms

*In the OpenClaw scenario, an attacker with Domain Admin privileges (obtained via the AD progression described in section 2) and collected integration secrets (cloud tokens, API keys) can potentially target **four classes of backups**:*

Class 1 — Local Volume Shadow Copies (VSS)

- **Mechanism**: deletion of shadow copies via native Windows tools (vssadmin.exe, wmic, PowerShell). As noted in the course material, defenders should monitor *"abnormal use of legitimate Windows tools such as vssadmin.exe to delete shadow copies, bcdedit.exe or wbadmin.exe to inhibit system recovery."*
- **Prerequisite**: local administrator rights (typically inherited from Domain Admin)
- **Detection**: SIEM monitoring of VSS deletions, execution restrictions on snapshot administration commands
- **Control**: copies out of AD account reach

Class 2 — Network Share Backups (NAS via SMB/CIFS)

- **Mechanism**: encryption or deletion of backup files accessible via network shares

- **Prerequisite**: credentials with write access to shares (typically Domain Admin or service accounts)
- **Detection**: monitoring of massive write access on backup shares
- **Control**: access segmentation (dedicated accounts outside AD), air-gapped or immutable backups

Class 3 — Dedicated Backup Infrastructure (Veeam, Commvault, etc.)

- **Mechanism**: deletion of jobs and restoration points via available administration interfaces (console, REST API, PowerShell)
- **Prerequisite**: access to backup software administration console — often accessible via the same AD accounts if no tiering is in place
- **Control**: network isolation of backup infrastructure, MFA on administration consoles, backup administration accounts **separate** from AD accounts, immutable backups (immutable flag at the storage level)

Class 4 — Cloud Backups

- **Mechanism**: revocation or rotation of cloud access tokens, deletion of snapshots/backups via cloud APIs
- **Prerequisite**: compromised cloud tokens or API keys (recovered from .env files or environment variables) — **it is not the Domain Admin privilege that grants this power, but separately recovered cloud secrets**
- **Control**: separation of cloud and AD credentials, MFA on cloud accounts, immutable retention policies on the cloud provider side, monitoring of deletion operations via cloud APIs

Table — Backup Classes: Neutralization Mechanisms and Defensive Controls

Backup Class	Neutralization Mechanism (generic)	Attacker Prerequisite	MITRE Technique	Defensive Control
Local volume snapshots (Volume Shadow Copies)	Deletion of restore points via native administration tools (Living-off-the-Land)	Local administrator or Domain Admin privileges	**T1490 — Inhibit System Recovery**	Monitoring of VSS deletions (SIEM alerts), execution restrictions for snapshot administration commands, copies outside AD account reach
Backups on network shares	Encryption or deletion of backup files	Credentials with write rights on shares (typically Domain	**T1486 — Data Encrypted**	Segmentation of backup share access (dedicated accounts outside AD), air-

(NAS via SMB/CIFS)	accessible via network shares	Admin or service accounts)	**for Impact + T1490**	gapped or immutable backups, monitoring of massive write access on backup shares
Dedicated backup infrastructure (Veeam, Commvault, etc.)	Deletion of jobs and restore points via available administration interfaces (console, REST API, PowerShell — depending on product version and configuration)	Access to backup solution administration console (admin credentials or network access to management interface)	**T1490**	Network isolation of backup infrastructure, MFA on administration consoles, backup admin accounts separate from AD accounts, immutable backup with verified retention
Cloud backups	Revocation or rotation of cloud access tokens, deletion of snapshots/backups via cloud APIs	Compromised cloud tokens or API keys (e.g. recovered from configuration files or environment variables) — it is not Domain Admin privilege that grants this power but separate cloud credentials	**T1490 + T1528 (Steal Application Access Token)**	Separation of cloud and AD credentials, MFA on cloud accounts, immutable retention policies on cloud provider side, monitoring of deletion operations on cloud backup APIs

5.3 AI Agent Amplification

The agentic amplification of this scenario lies in the ability of a compromised agent to quickly plan and execute a coordinated sequence of neutralization actions (network shares, dedicated backup infrastructure, local snapshots, cloud backups) in parallel, reducing the time window available for the defender to detect and respond.

This acceleration remains conditioned by the same factors as lateral movement (section 2.1):

- ***Effective permissions: deletion/encryption actions are only possible within the limits of the rights the agent has (or the credentials it has collected).***
- ***Planning capability and error robustness: the agent can potentially iterate on execution feedback (failed command → alternative attempt), which increases the reliability of multi-step sequences compared to a static script — but remains limited by the model's reasoning robustness.***
- ***Detection controls and operational limitations: guardrails such as action quotas, execution budgets, mandatory human validation for destructive actions, and per-tool restrictions reduce the blast radius — that is, the maximum damage scope of a single compromised agent.***

The impact of a successful injection strongly depends on the degree of agency (connected tools and authorized actions), which makes it a central argument for the least privilege principle applied to AI agents: every non-strictly-necessary tool and permission constitutes an additional attack surface.

5.4 Defensive Recommendations: The 3-2-1-1-0 Rule

Protection against backup neutralization relies on defense in depth applied to the backups themselves, as attackers frequently seek to delete or corrupt them before encrypting the IS. The historical 3-2-1 rule (3 copies, 2 media, 1 offsite) is now considered a necessary but insufficient baseline. ANSSI and industry best practices recommend the enhanced 3-2-1-1-0 rule, which adds an immutable or offline copy and regular restoration testing.

Concretely, 3-2-1-1-0 means: three copies of data (production + at least two backups), stored on two different media/technologies, with one offsite copy to withstand local disasters, plus one immutable or offline copy (inaccessible from the production network, even with Domain Admin privileges), and zero unverified backups.

The "0" is often the missing piece in practice: it requires verifying backup integrity and performing regular restoration tests (automated if possible), because an untested backup often equates to a useless backup in crisis situations.

Finally, to break the "Domain Admin → backup destruction" scenario, isolation is decisive: removing the backup infrastructure from the AD authentication and administration perimeter when possible (dedicated accounts, isolated credential vault), with MFA on backup management consoles.

5.5 Key Impact Data

Failure to protect backups has direct consequences on post-incident decisions:

- 49% of victims with encrypted data in 2025 paid the ransom to regain access Cyberlab
- 38% of organizations that paid more than the initial demand cited the fact that their backups had failed or were malfunctioning Cyberlab
- The median ransom payment fell to $1 million in 2025 (down from $2 million in 2024) Sophos, but for mid-size pharmaceutical companies the amount is calibrated to revenue
- Organizations with immutable backup infrastructure and regularly tested restores saw significantly lower rates of ransom payment and downtime, even when infected Veeam

Section conclusion: in the OpenClaw scenario, backup neutralization between D+4 and D+5 is the **prerequisite** for Phase 5 success (PromptLock deployment). Without this step, the organization could restore without paying. Defense rests on a simple principle: **separate the backup plane from the destruction plane** by isolating backups from the compromised AD perimeter.

6. MITRE ATT&CK / ATLAS Mapping — Phase 4

The table below maps Phase 4 techniques and tactics according to MITRE ATT&CK v15 and MITRE ATLAS. Identifiers are verified against primary sources; ATLAS tactics are qualified as such when they do not correspond to traditional Enterprise ATT&CK techniques.

Table — Phase 4 Matrix: Lateral Movement, Exfiltration and Backup Neutralization

Tactic	Technique	ID	Description (non-operational level)	Mapping Note
Execution	Command and Scripting Interpreter: PowerShell	**T1059.001**	Internal reconnaissance via native administration tools (LotL paradigm)	Direct mapping
Credential Access	OS Credential Dumping: LSASS Memory	**T1003.001**	In-memory credential extraction from authentication processes	Direct mapping. No offensive tool name — the technique describes the objective, not the implementation
Lateral Movement	Use Alternate Authentication Material: Pass-the-Hash	**T1550.002**	Reuse of hashes for authentication on other domain systems	Direct mapping
Lateral Movement	Remote Services: SMB/Windows Admin Shares	**T1021.002**	Propagation via administrative shares (ADMIN$, C$)	*Added — complements T1550.002 for the effective movement mechanism*
Credential Access	OS Credential Dumping: DCSync	**T1003.006**	Abuse of AD replication protocol to obtain directory secrets	Direct mapping
Persistence	Steal or Forge Kerberos Tickets: Golden Ticket	**T1558.001**	TGT forging for persistent domain access	Direct mapping. Access obtained depends on tiering policies — "unlimited access" is only true in the absence of privilege segmentation
ATLAS technique	LLM Prompt Injection	**AML.T0051**	Hijacking of OpenClaw agent via malicious content ingested from messaging channels	*AML.T0051 without sub-technique .001 for lack of confirmed primary ATLAS source for this ID. Direct/indirect distinction qualified in description [25]*
OWASP Agentic	Agent Goal Hijacking	**ASI01**	Compromised agent executes network actions in accordance with attacker objectives,	*OWASP Top 10 for Agentic Applications 2026 category, not a MITRE technique.*

			exploiting its legitimate tools and permissions	*Retained for descriptive relevance [160]*
ATLAS technique	Poison Training Data	**AML.T0020**	Compromise of internal chatbot model via targeted weight editing (ROME/PoisonGPT type). AML.T0020 covers poisoning at the training data/finetuning level — targeted weight editing is a variant	*If the attacker replaces a pre-trained model (asset substitution) rather than retraining on poisoned data, AML.T0020 is a proxy — ATLAS does not have a specific technique for post-training weight editing*
Exfiltration	Exfiltration Over C2 Channel	**T1041**	Exfiltration via HTTPS requests from the malicious skill to C2 infrastructure, camouflaged in agent outgoing traffic	Direct mapping for the skill/agent channel. The chatbot channel (if outgoing connector available) constitutes a separate exfiltration vector — to be mapped according to actual channel (T1041 if C2)
Exfiltration	*(Chatbot channel — conditional)*	**T1041 or T1048**	Exfiltration via application connector of poisoned chatbot, if it has an outgoing capability (webhook, API, external logging)	*Separate channel from the previous one. Active exfiltration requires an outgoing application capability — without this condition, only passive leakage is possible (cf. section 4.3)*
Impact	Inhibit System Recovery	**T1490**	Deletion of local volume snapshots (VSS), neutralization of dedicated backup infrastructure, encryption/deletion of backup files on network shares	Direct mapping for neutralization of recovery capabilities
Impact	Data Encrypted for Impact	**T1486**	Encryption of backup files accessible via network shares (NAS/SMB)	*Complements T1490 — encryption of backup data falls under T1486, deletion of recovery mechanisms under T1490*

Credential Access	Steal Application Access Token	**T1528**	Revocation/abuse of recovered cloud tokens to neutralize cloud backups	*Separated from T1490: cloud token revocation is not system recovery inhibition in the T1490 sense, but application token abuse enabling access to backup management APIs*

7. Synthesis: Operational State at D+5

In the OpenClaw scenario, at the end of Phase 4, the attacker potentially has several complementary capabilities on PharmEurys SA's information system. The table below summarizes the state of each capability with an evidence level, conditional factors, and main defensive fragility.

Table — State of Offensive Capabilities at D+5

Capability	Vector	OpenClaw's Role	Status D+5	Detectability	Maintenance Condition / Fragility
Privileged AD access	Golden Ticket (T1558.001)	Exposed instance identified in Phase 1 → initial access → AD progression	Active, persistent as long as KRBTGT secret is not sanitized (double rotation)	Detection possible via abnormal Kerberos ticket monitoring (lifetime, SID, encryption type), replication request alerts	**Fragility: a double KRBTGT rotation invalidates the Golden Ticket. Persistence depends on the absence of this remediation operation**
Lateral movement via Slack agent	Indirect prompt injection (AML.T0051, ASI01)	Hijacked agent executing actions via terminal and legitimate tools	Active, weak signal if actions use legitimate tools with user permissions	Behavioral detection (tool usage anomalies, volumes, schedules), tool call monitoring, file access → outgoing request correlation	**Fragility: tool allowlist, human confirmation, sandboxing, agent permission revocation**

AI chatbot backdoor	Modified model (ROME/PoisonGPT type)	Access to chatbot server via elevated privileges obtained in Phase 4	Active, discreet — detection difficult without dedicated controls (performance deviation ~0.1% on standard benchmarks in PoisonGPT demo)	Deployed model audit (hash/provenance verification), targeted evaluation on known triggers, abnormal response monitoring	**Fragility: model integrity verification (cryptographic hash), signed provenance, redeployment from trusted source**
R&D exfiltration	Agent channel (skill HTTPS, T1041) + chatbot channel (conditional)	Orchestration and execution on agent side, HTTPS traffic camouflaged in normal API traffic	Data potentially exfiltrated, if outgoing channels are operational and not filtered by allowlist/DLP	Egress control (destination allowlist), DLP, volumetric analysis, correlation of sensitive data access → outgoing requests	**Fragility: strict egress allowlist, TLS inspection, DLP on outgoing content**
Neutralized backups	LotL + AD privileges + cloud secrets (T1490, T1486, T1528)	Cloud tokens recovered via agent configuration files	Recovery capabilities inhibited / restoration significantly compromised	Alerts on VSS deletions, monitoring of deletion operations on backup APIs, audit of backup administration console access	**Fragility: immutable backups, air-gapped copies outside AD perimeter, isolated backup accounts, 3-2-1-1-0 rule**

References

Note: Numbering [111] to [145], continuing from Phases 1–3 ([1]–[110]).

[111] Lockheed Martin, « Cyber Kill Chain Framework — C2, Lateral Movement, Actions on Objectives ». https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[112] CrowdStrike, « Indirect Prompt Injection Attacks: Hidden AI Risks » (mouvement latéral via agents compromis), décembre 2025. https://www.crowdstrike.com/en-us/blog/indirect-prompt-injection-attacks-hidden-ai-risks/

[113] TechTarget / SearchSecurity, « News brief: AI threats to shape 2026 cybersecurity ». M. Sachs (CIS), J. Grady (Omdia), P. Harrington (Forrester). https://www.techtarget.com/searchsecurity/news/366637045/

[114] MITRE ATT&CK, « T1059 Command and Scripting Interpreter » et Living-off-the-Land Binaries, v15. https://attack.mitre.org/

[115] Verizon, « 2025 Data Breach Investigations Report » (DBIR). 74 % violations impliquent identités compromises. Kill chain AD typique.

[116] CIS, « Mimikatz: The Finest in Post-Exploitation » (sekurlsa, lsadump, DCSync, Golden Ticket). https://www.cisecurity.org/insights/blog/mimikatz-the-finest-in-post-exploitation

[117] S. Metcalf (ADSecurity.org), « Attack Methods for Gaining Domain Admin Rights in Active Directory ». Credential Theft Shuffle. https://adsecurity.org/?p=2362

[118] Netwrix, « DCSync Attack Using Mimikatz Detection ». https://www.netwrix.com/privilege_escalation_using_mimikatz_dcsync.html

[119] Stellar Cyber, « Top 10 Agentic SOC Platforms for 2026 ». Cas Change Healthcare (190M patients, 9 jours, credential unique). https://stellarcyber.ai/learn/top-10-agentic-soc-platforms/

[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/

[121] InstaTunnel, « Prompt-to-Insider Threat: When AI Agents Become Double Agents ». CVE-2025-32711 EchoLeak (M365 Copilot, CVSS 9.3), février 2026. https://instatunnel.my/blog/prompt-to-insider-threat/

[122] HackerNoob / Information 2026, 17(1), 54, « Prompt Injection Attacks in LLM and AI Agent Systems: A Comprehensive Review » (taux succès 66,9–84,1 %, 73 % déploiements affectés). doi:10.3390/info17010054

[123] Mithril Security, « PoisonGPT: How to poison LLM supply chain on Hugging Face » (ROME, GPT-J-6B, Δ 0,1 %). https://blog.mithrilsecurity.io/poisongpt/

[124] Barracuda Networks, « PoisonGPT: Weaponizing AI for disinformation », sept. 2025. https://blog.barracuda.com/2025/09/11/poisongpt-weaponizing-ai-disinformation

[125] Anthropic, « Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training », 2024. ArXiv.

[126] Phase 3, [80]–[82]. Exfiltration via skill OpenClaw : curl C2 encapsulé dans trafic HTTPS légitime du gateway.

[127] S. Willison, « AI agents have a lethal trifecta of risks » (private data + untrusted content + external communication). Réf. Phase 3 [90].

[128] Sophos, « The State of Ransomware 2025 ». 94 % attaques ciblent sauvegardes, 57 % réussissent.

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

[130] Microsoft, « Guidance to mitigate critical threats to AD Domain Services in 2025 ». https://www.microsoft.com/en-us/windows-server/blog/2025/12/09/

[131] Control Risks, « The Agentic Shift: How Autonomous AI Is Reshaping the Global Threat Landscape ». https://www.controlrisks.com/our-thinking/insights/the-agentic-shift

[132] Obsidian Security, « Prompt Injection Attacks: The Most Common AI Exploit in 2025 » (privilèges excessifs SaaS, mouvement latéral). https://www.obsidiansecurity.com/blog/prompt-injection

[133] Lakera, « The Year of the Agent: Q4 2025 Attacks » (attaques indirectes < tentatives que directes, system prompt extraction). https://www.lakera.ai/blog/the-year-of-the-agent

[134] Sombrainc, « LLM Security Risks in 2026 » (ServiceNow second-order injection, agent privilege escalation). https://sombrainc.com/blog/llm-security-risks-2026

[135] OpenAI, « Understanding prompt injections: a frontier security challenge », janvier 2026. https://openai.com/index/prompt-injections/

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*

[25] MITRE ATT&CK, « Active Scanning: Vulnerability Scanning », Sub-technique T1595.002. https://attack.mitre.org/techniques/T1595/002/

→ *Defined in Phase 1*

[27] The Shadowserver Foundation, « CRITICAL: Vulnerable HTTP Report ». https://www.shadowserver.org/what-we-do/network-reporting/vulnerable-http-report/

→ *Defined in Phase 1*

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

→ *Defined in Phase 5*

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

→ *Defined in Phase 5*