
TECHNICAL REPORT — TR-2026-01

Operation "OpenClaw"

Anatomy of an AI-Driven Cyberattack Against a Pharmaceutical Company

Phase 3 — Installation and Execution

Malicious Skill, Credential Theft, Deepfake

and VPN Exploitation — From Delivery to Initial Intrusion (D-7 to D-Day)

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 the third phase of a fictional but realistic attack scenario. It details the installation and execution mechanisms of three independent initial access vectors: malicious skill supply chain, AI agent credential theft via infostealer, and VPN exploitation.

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 third installment of the Operation "OpenClaw" analysis and covers Phase 3 — Delivery and initial access (D−7 to D-Day), during which the offensive artifacts prepared in Phase 2 are deployed against PharmEurys SA. Three independent and complementary initial access vectors are analyzed: (1) installation of the malicious skill via the ClawHub supply chain, (2) theft of agent identity credentials via infostealer (Vidar), and (3) exploitation of the Fortinet VPN vulnerability (CVE-2024-55591).

This document analyzes the success conditions of each vector and the defensive controls enabling kill chain interruption at each stage. It does not describe reproducible intrusion procedures; technical details remain at the conceptual level required for risk analysis and control derivation.

Keywords: initial access, malicious skill, credential theft, AI agent, supply chain, VPN exploitation, Fortinet, MITRE T1078, T1195.002, AML.T0051, OWASP LLM01:2025

1. Introduction: From Weaponization to Initial Access

Phase 1 of Operation OpenClaw (D−30 to D−15) produced actionable intelligence on PharmEurys SA: an exposed OpenClaw instance identified via Internet asset databases, a functional organizational chart reconstructed via social graph mining, and the inference of a potential Fortinet VPN vulnerability (CVE-2024-55591). Phase 2 (D−15 to D−7) transformed this intelligence into three families of offensive artifacts: a malicious supply chain skill, the PromptLock LLM-driven ransomware engine, and indirect prompt injection payloads.

Phase 3 — Delivery and initial access — corresponds, in the Lockheed Martin Cyber Kill Chain, to the stages of delivery (Delivery), exploitation (Exploitation) and installation (Installation): this is the moment when offensive artifacts make contact with the target and establish the first footholds in the information system.

In the context of AI agents, the attack surface is multi-channel: extensions/skills, ingested content, connectors, identities and network access. The attacker does not depend on a single vector but orchestrates multiple complementary vectors, increasing the probability of at least one succeeding.

Willison's lethal trifecta — access to private data, exposure to untrusted content, and external communication capability [127] — provides a useful criterion for identifying configurations at risk. In the case of OpenClaw, the simultaneous presence of connectors (Slack, Outlook, terminal), external ingested sources (web pages, documents), and network capabilities (API calls, curl) creates a configuration that meets all three properties.

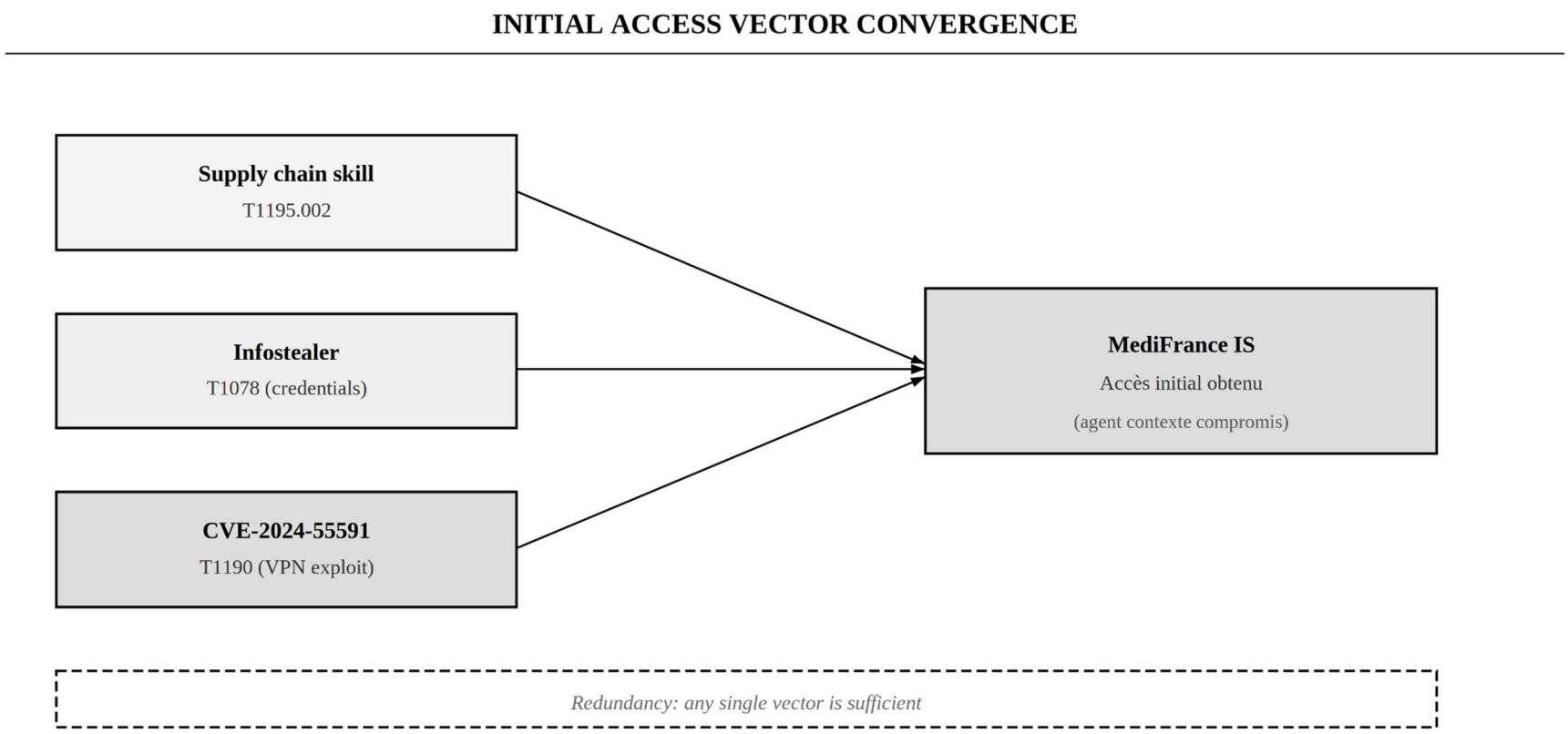

Figure 10. Convergence of the three initial access vectors toward the MediFrance IS. Each vector operates independently (skill supply chain, infostealer, VPN exploitation). This strategic redundancy increases the probability of successful initial access, even if one vector is blocked by defenses.

Phase 3 Operational Objectives

- Deliver and install the "PharmaResearch Assistant" skill on the target OpenClaw instance
- Exfiltrate agent identity credentials via commodity infostealer
- Exploit VPN vulnerability CVE-2024-55591 for parallel network access
- Establish three independent and complementary footholds in the target information system

2. Malicious Skill Delivery

2.1 Installation Vector: The Supply Chain as a Weapon

1) Snyk ToxicSkills figures: overall OK, but refine the formulation

- Snyk reports that 13.4% of skills (534) contain at least one "critical-level" issue, and that in total 36.82% (1,467) have at least one flaw.
- Snyk also reports that 91% of confirmed malicious skills combine traditional malware and prompt injection.

2) CVE-2026-25253: the vector is not "in the skill config"

According to the advisory (GitLab, NVD, reprints), the core of CVE-2026-25253 is that the control UI trusts the gatewayUrl parameter in the query string (URL), auto-connects via WebSocket, and sends the authentication token to the attacker's server. It is NOT a flaw in a SKILL.md or in the skill configuration itself.

3) "Remote code execution in one click... without further interaction"

"The 'PharmaResearch Assistant' skill, published on ClawHub during Phase 2, is discovered by an R&D employee while searching for monitoring skills. The installation vector relies on the default ClawHub configuration: installation requires only a user action (adding the skill to the agent configuration), without systematic security review by the registry."

"Furthermore, vulnerability CVE-2026-25253 (CVSS 8.8) illustrates a 'one-click' vector where the control interface trusts the gatewayUrl parameter provided in the URL and automatically initiates a WebSocket connection, sending the authentication token to the attacker's server. The resulting impact — remote code execution on the host machine — passes through abuse of the exfiltrated token and access to the victim's local gateway."

2.2 Silent Execution Mechanism

In the OpenClaw scenario, the malicious skill prepared in Phase 2 is published on the ClawHub registry and discovered by a PharmEurys SA R&D researcher during a search for pharmaceutical monitoring skills. The installation follows the standard ClawHub workflow: the user adds the skill to their agent configuration, granting SKILL.md instructions implicit trust at the level of the agent's execution context.

Scale of the Attack Surface (Snyk ToxicSkills, February 2026)

The Snyk "ToxicSkills" study highlights the systemic scale of this attack surface: of 3,984 ClawHub skills analyzed, 534 (13.4%) present at least one critical-level security finding, and 1,467 (36.82%) have at least one vulnerability. Among confirmed malicious skills, 91% combine at least one prompt injection technique with a traditional malware component (dropper, infostealer, backdoor).

CVE-2026-25253: One-Click Token Exfiltration Vector

Vulnerability CVE-2026-25253 (CVSS 8.8) illustrates a complementary initial access vector, disclosed in February 2026 [7][7]. The exploited mechanism is as follows: the Control Panel UI trusts the gatewayUrl parameter provided in the URL, automatically initiates a WebSocket connection to the attacker-specified address, and sends the authentication token. This one-click chain allows the attacker to obtain the gateway token, connect to the victim's local gateway, and execute privileged actions — potentially including remote code execution on the host machine.

Clarification on required interaction: this vector requires the victim to visit or click a link containing the malicious gatewayUrl parameter — it is therefore a one-click attack and not a zero-click attack. The "one-click RCE" qualification refers to the fact that a single user action triggers the complete exploitation chain without further interaction.

In the OpenClaw scenario, this vulnerability constitutes an alternative initial access vector to supply chain installation: if the targeted R&D employee does not discover the malicious skill on ClawHub, the attacker can send them a link containing the malicious gatewayUrl via email or Slack.

2.3 Taxonomy of Skill-Based Installation Vectors

The following table synthesizes the installation vectors identified in the OpenClaw ecosystem, with impact bounded by the agent's effective permissions and an explicit evidence level for each vector.

Table — Installation Vectors: Techniques, Bounded Impacts and Sources

Vector	Technique	Impact (bounded)	Evidence Level	Mapping Note
Malicious skill on community registry	Supply chain via trusted extension registry (T1195.002)	Action execution within the agent's permission scope — tools, connectors, egress. Actual impact depends on configuration: sandboxing, allowlist, egress policy	Documented (Koi Security, Snyk ToxicSkills, Cisco AI Defense)	Direct mapping T1195.002
CVE-2026-25253	Unvalidated gatewayUrl parameter in URL → automatic WebSocket connection → gateway token	One-click chain: attacker obtains authentication token, connects to victim's local gateway, and can execute privileged actions —	Documented (advisory GitLab, NVD, technical analyses runZero, SecurityScorecard)	*RCE impact goes through token abuse and gateway access, not*

	exfiltration to attacker infrastructure	potentially including RCE on host machine		*direct browser execution*
Injection via HEARTBEAT.md	Indirect prompt injection (via web page, document) causing the agent to write instructions in HEARTBEAT.md	Cross-session persistence: OpenClaw's system prompt includes HEARTBEAT.md content when present, integrating malicious instructions into the agent's context at every subsequent launch	Documented (HiddenLayer, consistent with OpenClaw documentation — heartbeat prompt explicitly reads this file)	Corresponds to stage 4 (Persistence) of the Promptware Kill Chain [120]
Ranking manipulation	Manipulation of registry popularity and reputation signals (artificial download metric inflation, fraudulent reviews)	Increased visibility of malicious skill, maximizing probability of discovery and installation by legitimate users [11]	Documented (Dvuln — experimental demonstration, AuthMind/iKangai reprints)	*T1608 (Stage Capabilities) at parent level. T1608.006 (SEO Poisoning) specifically targets search engines and does not apply to internal marketplace ranking*
Injection via logs (log poisoning)	Injection of malicious content into system or application logs subsequently ingested by an agent (e.g. automatic log summarization, RAG pipeline fed by log aggregation)	Alteration of decisions or tool behaviors if logs are processed as trusted source by the agent	**Hypothetical vector — this scenario is technically plausible in architectures where an agent ingests logs as data source, but is not specifically documented in the OpenClaw context**	—

Notes:

- **Impact bounded by permissions: each vector produces effects limited by the effective agent configuration (sandboxing, tool allowlist, egress policy, containerization). The "worst case" impact assumes an insufficiently restricted configuration, documented as prevalent in initial deployments.**

- **Vector complementarity: in the OpenClaw scenario, these vectors are not mutually exclusive. Supply chain installation (malicious skill) and CVE-2026-25253 exploitation (one-click token exfiltration) can be combined to maximize the probability of initial access.**

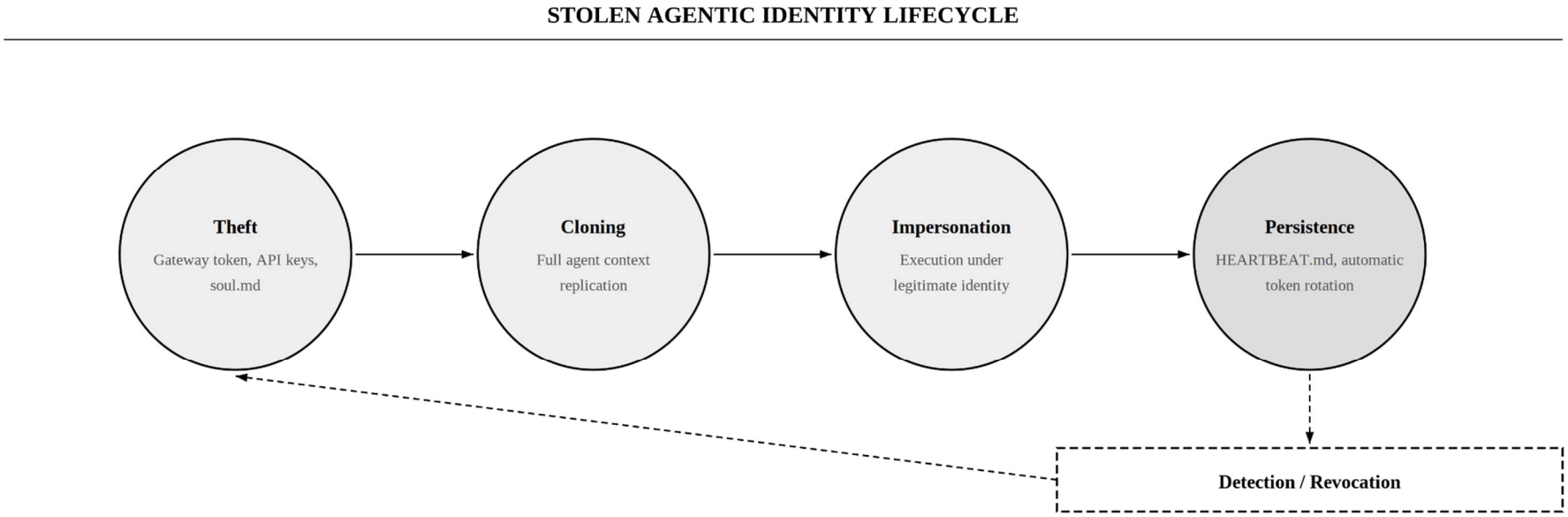

Figure 11. Stolen agentic identity lifecycle. From initial theft (tokens, keys, configuration) to agent context cloning, then identity impersonation and persistence via HEARTBEAT.md. The detection/revocation loop (dashed) represents the defensive mechanism for breaking the cycle.

3. Credential Theft via Infostealer

3.1 Anatomy of AI Agent Identity Exfiltration

A critical vector in Phase 3 is the theft of an OpenClaw agent's identity artifacts via an infostealer. In February 2026, Hudson Rock publicly documented one of the first cases of large-scale exfiltration of AI agent configuration files by commodity infostealers.

The exfiltrated files documented by primary sources include three main artifacts: (i) openclaw.json, containing the gateway authentication token, associated email address, and workspace path; (ii) device.json, containing cryptographic keys used for secure pairing and communication signing; and (iii) soul.md, containing user-configured operational principles, behavioral directives, and ethical limits.

The combined compromise of these artifacts — access token, cryptographic keys, behavioral identity and personal context — potentially enables, depending on deployment context and controls in place, an attacker to impersonate the agent, access its integrations, and manipulate its behavior.

3.2 Consequences of Theft: Agentic Identity Cloning

The possession of OpenClaw identity artifacts (gateway token, pairing/signing keys, behavioral directives) can enable agent impersonation on another machine, creating a "shadow agent" with the same access and behavioral profile as the legitimate agent.

Authentication Mechanisms and Exploitation Conditions

OpenClaw gateway authentication mechanisms rely on tokens and scoped device tokens issued after pairing. Compromise of these elements can result in unauthorized access, but the actual impact depends on several factors:

- **Token validity: if the gateway implements periodic token rotation or revocation upon anomaly detection, the exploitation window is limited.**
- **Device token scoping: permissions associated with the device token may restrict available actions for the attacker depending on deployment configuration.**
- **Access and tool policies: tool allowlists and egress restrictions configured on the gateway side also apply to the shadow agent.**

The impersonation is therefore not an automatic "full clone," but an authenticated access whose scope depends on controls in place. In the absence of token rotation and multi-session monitoring, the exploitation window can be significant.

Secret Storage and Theft Surface

OpenClaw project security documentation explicitly states that the contents of the configuration directory (~/.openclaw/) must be considered sensitive and recommends hardening file permissions (600/700 modes). However, by default, these files are stored in plain text without encryption at rest.

Observed Scale

Public data reported by Hudson Rock indicates that Vidar-type infostealers exfiltrated OpenClaw configuration files from a significant number of infected machines in early 2026. This observation confirms that AI agent identity artifacts are now part of the standard loot of commodity infostealers.

3.3 Mapping of Exfiltrated Identity Artifacts

Table — Identity Artifacts: Documented Content, Bounded Impact and MITRE Mapping

File	Content (source: public analyses Hudson Rock / TheHackerNews)	Impact (conditional)	MITRE ATT&CK
openclaw.json	Gateway access token(s), associated email address, workspace path [13]	Possible impersonation of application access to gateway as long as tokens remain valid and unrevoked	**T1528 — Steal Application Access Token**
device.json	Cryptographic keys used for secure pairing and communication signing [13]	Reinforces impersonation capability if these keys are used in the gateway authentication process	**T1528 — Steal Application Access Token**
soul.md	Operational principles, behavioral directives, ethical limits configured by user [13]	Intelligence on agent behavioral profile; possibility of behavior manipulation and functional persistence	**N/A — no direct ATT&CK mapping (AI agent-specific local state file, outside scope of existing techniques)**
.env (adjacent risk)	Environment variables, third-party API keys, access tokens for integrated services	Potential pivot to third-party services depending on key scopes present	**T1552.001 — Credentials in Files**

Notes on mapping:

- **T1528 (Steal Application Access Token) replaces T1078.004 (Cloud Accounts) for openclaw.json. T1078.004 targets abuse of cloud accounts (SaaS/IaaS identities); an agent gateway token more closely corresponds to an application access token in the T1528 sense.**
- **T1213 (Data from Information Repositories) is not applicable to soul.md: T1213 targets systems like SharePoint/Confluence/wiki, not a local agent memory file. In the absence of a specific ATT&CK technique covering local state files of AI agents, this artifact is classified as N/A.**
- **.env is included as an adjacent risk: Vidar-type infostealers typically target environment files, but .env exfiltration is not explicitly documented in the Hudson Rock case. Inclusion is based on the known behavioral profile of this infostealer family.**

Impact of credential theft — bounded analysis

The theft of these artifacts increases the risk of agent impersonation and unauthorized application access, particularly when access tokens and integration secrets are present on disk without encryption at rest. However, the actual impact depends on the security controls in place:

- **Data access: the attacker can potentially access data to which the agent is connected (documents, messaging, cloud services) — within the limits of scopes and permissions effectively granted to the compromised token.**
- **Execution via agent tools: the attacker can potentially execute actions via installed skills and authorized tools — within the limits of allowlists and tool policies configured on the gateway.**
- **Detectability: shadow agent activity uses the same channels as the legitimate agent, which complicates detection — without making it impossible. Detection controls remain operational: simultaneous session monitoring, source IP correlation, abnormal volume analysis.**

4. Fortinet VPN Exploitation

4.1 Parallel Network Access Vector

In parallel with AI agent-oriented initial access vectors (skill supply chain, credential theft), the attacker exploits the Fortinet VPN vulnerability identified during the Phase 1 reconnaissance. This vector provides direct network access to PharmEurys 's internal infrastructure, independent of agent-specific vectors.

Vulnerability CVE-2024-55591 (CVSS 9.6) is a logic flaw of the Authentication Bypass type (CWE-288 — Authentication Bypass Using an Alternate Channel) residing in the WebSocket module (jsconsole) of the FortiOS management interface. The exploitation mechanism operates in three stages:

- An unauthenticated remote attacker can manipulate WebSocket requests to force the management API to deliver a valid "Super Admin" session context, bypassing all normal access controls.
- The attack requires no memory corruption (no buffer overflow) and leaves few traces in standard system logs, as the generated session appears legitimate to the audit subsystem.
- The access obtained is at the System/Root level, allowing immediate creation of persistent users or log disabling.

This CVE was actively exploited in the wild in early 2025, documented by Fortinet PSIRT (FG-IR-24-535), watchTowr Labs, Tenable Research, and ANSSI (CERTFR-2025-ALE-002). It is listed in the CISA KEV catalog.

Reminder: the inference of PharmEurys *'s exposure to this CVE relies on the passive versioning correlation performed in Phase 1 (Shodan/Censys), with a confidence level conditional on fingerprint quality (cf. Phase 1, section 3.1).*

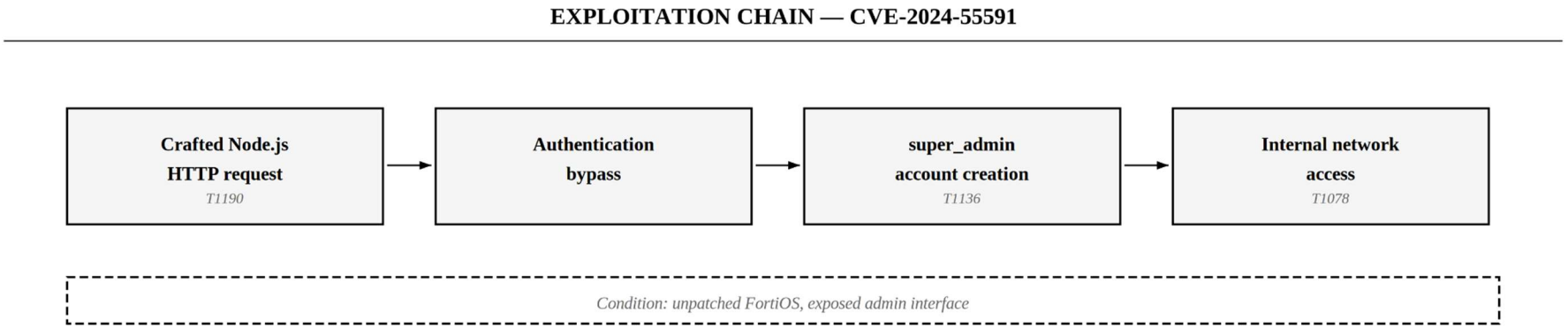

Figure 12. CVE-2024-55591 exploitation chain. Four-step sequence from crafted HTTP request to internal network access. Exploitation is conditional on unpatched FortiOS.

4.2 Exploitation Methodology

Exploitation of CVE-2024-55591 requires that the FortiOS administration portal (jsconsole module) be accessible from the Internet and that the firmware version predate the January 2025 patches. In the OpenClaw scenario, these conditions were inferred during the Phase 1 passive reconnaissance.

Table — Success Conditions and Defensive Controls (CVE-2024-55591)

Condition	Description	Defensive Control
Exposed administration portal	The jsconsole module is accessible from the Internet (port 443)	Restrict administration access to internal networks, ACL on management interface
Vulnerable firmware version	FortiOS < 7.0.17 (vulnerable ranges documented by Fortinet PSIRT)	Patch application, accelerated patching policy for perimeter equipment
Absence of admin session monitoring	The Super Admin session created by the exploit appears legitimate	Monitoring of admin account creations, alerts on admin sessions outside normal scope, SIEM correlation
Post-exploitation persistence	Super Admin access allows creation of persistent users and configuration modification	Admin account audit, configuration integrity verification, regular config backup and comparison (ANSSI 3-2-1 rule) [158]

Note: the detailed exploitation methodology (WebSocket request sequence, payload structure) is not described — it is documented in the technical analyses by watchTowr Labs and Tenable Research. This document focuses on the success conditions and applicable defensive controls.

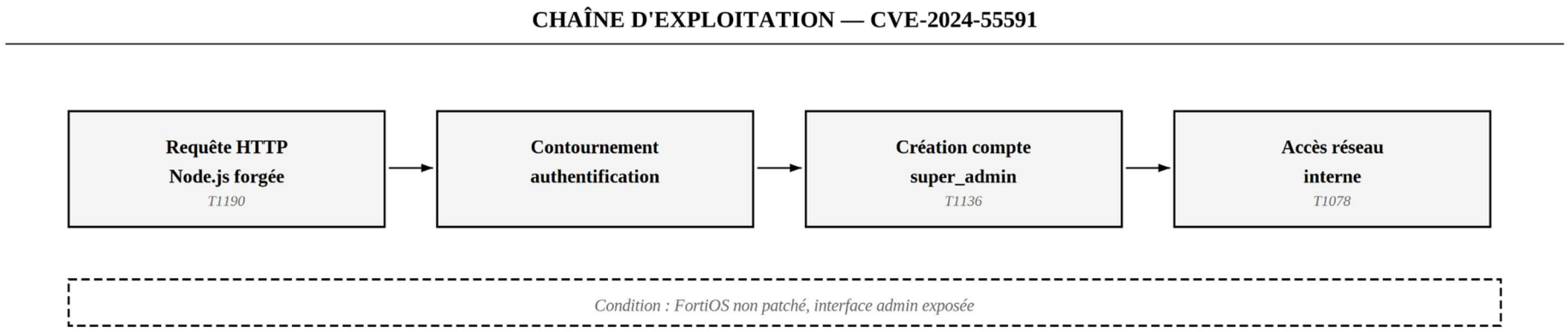

Figure 12. Chaîne d'exploitation CVE-2024-55591. Séquence en quatre étapes menant de la requête HTTP forgée à l'accès réseau interne. L'exploitation est conditionnelle au non-patching de FortiOS.

4.3 Strategic Redundancy

This VPN vector constitutes complementary access: if OpenClaw agent-related vectors are detected and remediated, network access via the compromised VPN can persist, offering a return path into the information system independent of agent-specific controls.

Detectability: although CVE-2024-55591 exploitation is stealthy at the FortiOS log level (apparently legitimate session), detection signals remain exploitable — monitoring of admin account creations, alerts on admin sessions outside normal scope, SIEM correlation on concurrent sessions.

5. MITRE ATT&CK / ATLAS Mapping

The table below maps Phase 3 techniques and tactics according to MITRE ATT&CK v15 and MITRE ATLAS. Identifiers are verified against primary sources; when no existing technique precisely covers the described behavior, this is explicitly noted.

Table — Phase 3 Matrix: Delivery and Initial Access

Tactic	Technique	ID	Description	Vector	Mapping Note
Initial Access	Compromise Software Supply Chain	**T1195.002**	Installation of a malicious skill from the ClawHub community registry, exploiting trust in the marketplace	Skill supply chain	Direct mapping [25]
Initial Access	Exploit Public-Facing Application	**T1190**	Exploitation of CVE-2024-55591 (Auth Bypass WebSocket/jsconsole, CVSS 9.6) on the exposed FortiOS portal	Fortinet VPN	Direct mapping [77]
Credential Access	Steal Application Access Token	**T1528**	Theft of gateway access tokens (openclaw.json) and pairing/signing keys (device.json) via Vidar infostealer	Infostealer	*Replaces T1078.004 (Cloud Accounts) which targets SaaS/IaaS identities. The gateway token is an application access token [13]*
Credential Access	Credentials in Files	**T1552.001**	Integration secrets stored on disk (.env, configuration files) without encryption at rest	Infostealer	*Replaces T1555 (Credentials from Password Stores) which targets password managers, not configuration files [13]*
Execution	LLM Prompt Injection	**AML.T0051 (ATLAS)**	Indirect injection via skill content (SKILL.md) and persistence via HEARTBEAT.md	Skill / Ingested content	*Without sub-technique .001 for lack of primary ATLAS source. Direct/indirect distinction qualified in description [120]*

Collection	*AI agent local state file*	**N/A**	Exfiltration of soul.md (operational principles, directives, memory)	Infostealer	*T1213 (Data from Information Repositories) targets SharePoint/Confluence-type repositories, not a local file. No ATT&CK technique specifically covering AI agent local state files*

6. Synthesis: State of Compromise at D-Day

6.1 Operational Dashboard

In the OpenClaw scenario, at the end of Phase 3, the attacker potentially has three independent and complementary access vectors to PharmEurys *SA's information system. The table below summarizes the state of each vector.*

Table — State of Access Vectors at D-Day

Vector	Evidence Level	Access / Impact (bounded)	Primary Detection Surface	Weakness / Mitigation
Malicious skill on registry	Prospective scenario (based on documented components: Koi Security, Cisco, Snyk)	Actions within agent permission scope — tools, connectors, egress. Scope depends on configuration: sandboxing, tool allowlist, egress policy	Agent telemetry (tool call monitoring), network egress control, endpoint detection (abnormal processes initiated by agent)	Tool and egress restrictions, skill sandboxing, code review, publisher cryptographic signing, allowlist [25]
Infostealer credentials (Vidar)	Observed in the wild (Hudson Rock, February 2026 — agent identity artifact theft documented) [13]	Possible agent impersonation via tokens as long as they remain valid and unrevoked. Access inherits the legitimate agent's application permissions	Gateway and SaaS logs (simultaneous sessions, unusual source IP, abnormal volumes), token abuse detection (T1528)	**Token rotation and revocation, abnormal usage detection, enhanced device binding, identity artifact encryption at rest [13]**
Fortinet VPN (CVE-2024-55591)	Prospective scenario (real CVE, exploited in the wild, PharmEurys exposure inferred by passive correlation with conditional confidence) [77]	Network access at the level of VPN-accessible segments — scope depends on network segmentation, VPN positioning in architecture	IDS/IPS, VPN logs (admin account creations, sessions outside normal scope), SIEM correlation	**Patch application (primary mitigation), restrict admin access to internal networks, network segmentation, admin account monitoring [1]**

6.2 Defensive Posture Analysis

The convergence of three independent vectors (AI supply chain, agent credential theft, perimeter exploitation) shifts detection toward several complementary layers:

- **Agent layer (specific to AI ecosystems): tool call monitoring, extension governance, destination-based egress control, persistent memory integrity. This is the least mature layer in most organizations today.**
- **Endpoint layer (EDR): infostealer detection, sensitive configuration file protection, abnormal process monitoring. EDR is an endpoint control, not a perimeter control — it complements but does not replace network controls.**
- **Network layer (perimeter): IDS/IPS, WAF, egress control, segmentation. Effectiveness is reduced when malicious traffic uses legitimate HTTPS channels (LLM API, cloud services), but remains essential for detecting lateral movement and anomalous exfiltration patterns.**
- **Identity/SaaS layer: gateway session monitoring, token abuse detection (T1528), rotation and revocation, session metadata correlation [13].**

None of these layers is sufficient in isolation. Defense in depth against a multi-vector strategy requires cross-layer correlation — which constitutes precisely the operational challenge for the security teams of targeted organizations.

6.3 Transition to Phase 4 — Lateral Movement and Exfiltration (D+1 to D+5)

The next phase will analyze post-initial-access propagation mechanisms in the PharmEurys environment: lateral movement via compromised agent connectors (stage 6 of the Promptware Kill Chain), Active Directory compromise, and preparation of data exfiltration.

6.4 Convergence of Attack Surfaces

Operation OpenClaw illustrates a convergence of attack surfaces where historically distinct paradigms — supply chain compromise (skill registry), agent identity theft (infostealer), and perimeter exploitation (VPN CVE) — combine to create a multi-dimensional threat that exceeds the detection capabilities of any single defense layer.

OWASP formalized these risks in its Top 10 for Agentic Applications 2026, which explicitly includes agentic supply chain vulnerabilities (ASI04 — Agentic Supply Chain Vulnerabilities) and uncontrolled autonomy as emerging risk categories.

Autonomous AI agents compose tools, identity, memory and content ingestion in a single execution environment, creating a multi-dimensional attack surface. Recent literature identifies three roles that the agent simultaneously plays in the attack chain:

- **Entry point (vector): the compromised skill or ingested content alters the agent's behavior (OWASP LLM01, Promptware Kill Chain — stages 1–2) [120].**
- **Surface to compromise: the agent's identity (tokens, keys, memory) is stolen or impersonated, offering authenticated access to its integrations (T1528) [13].**
- **Amplifier: the compromised agent plans and executes multi-step actions autonomously — internal reconnaissance, lateral movement via connectors, exfiltration — in accordance with stages 3–7 of the Promptware Kill Chain [120].**

This triple function — vector, surface, amplifier — is identified in several recent works on agentic risks and constitutes an emergent property of systems combining autonomous planning and privileged access to enterprise tools.

6.5 Lessons for Defense

Defense against an OpenClaw-type attack cannot rely on a single mechanism, as it combines extension supply chain compromise, semantic hijacking (prompt injection), identity abuse (token theft), and conventional perimeter exploitation. The following sections synthesize priority controls per attack surface.

Extension Governance and Tool Isolation

Cisco AI Threat Research emphasizes that the major risk from agents comes from the combination "prompt injection + tool access," and recommends a defense-in-depth approach based on tool isolation, guardrails, and execution monitoring.

Secret Protection and Token Governance

OpenClaw security documentation explicitly states that files under ~/.openclaw/ must be considered sensitive and recommends permission hardening (600/700 modes). However, by default, files are stored in plain text. Priority controls include:

- **Encryption at rest of configuration and secret files.**
- **Gateway token rotation and revocation — MITRE recommends these controls to counter application token abuse (T1528), including API call monitoring and session metadata correlation.**
- **Abnormal token usage detection: simultaneous sessions from different IP addresses or devices, access outside normal scope.**

Protection of local secrets constitutes the first line of defense against the infostealer vector, but it depends on host security controls (EDR, local privilege management, disk encryption).

Perimeter Vulnerability Management

The exploitation of CVE-2024-55591 is a reminder that agentic controls do not exempt from basic security hygiene: accelerated patching policy for exposed equipment, restriction of administration access to internal networks, network segmentation, and admin account monitoring.

Synthesis: Defensive Checklist by Attack Surface

Surface	Priority Controls
Skill supply chain	Code review / skill audit, publisher cryptographic signing, allowlisting, execution sandboxing
Prompt injection	Data/instruction separation, tool call monitoring, tool restrictions, persistent memory governance
Agent identity (tokens/secrets)	Encryption at rest, token rotation/revocation, T1528 abuse detection, local permission hardening

Network vulnerabilities	Accelerated patching, admin access restriction, segmentation, admin account and VPN session monitoring
Cross-layer observability	Cross-layer correlation (agent + endpoint + network + identity), behavioral anomaly alerts

→ Perspectives: Beyond D-Day

The three accesses potentially established at the end of Phase 3 constitute the foundations for subsequent Kill Chain phases. Post-intrusion tactics — lateral movement, privilege escalation, data exfiltration, ransomware deployment — will be analyzed in Phases 4 and 5.

References

Note: Numbering [76] to [110], continuing from Phases 1 ([1]–[40]) and 2 ([41]–[75]).

[76] Lockheed Martin, « Cyber Kill Chain Framework — Delivery, Exploitation, Installation ». https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

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

[78] S. Willison, « The Lethal Trifecta of AI agents » (accès données privées + contenu non approuvé + communication externe), réf. Palo Alto Networks Unit 42, février 2026.

[79] 1Password, « From magic to malware: How OpenClaw's agent skills become an attack surface », février 2026. https://1password.com/blog/from-magic-to-malware

[80] Snyk, « ToxicSkills: 3 984 skills auditées, 534 (13,4 %) problèmes critiques, 76 payloads malveillants confirmés. 91 % combinaient prompt injection et malware traditionnel », février 2026. Réf. Barrack.ai/TechInformed.

[81] SOCRadar, « CVE-2026-25253 (CVSS 8.8) : 1-click RCE via exfiltration token auth par paramètre gatewayUrl malveillant », février 2026. https://socradar.io/openclaw-rce-vulnerability/

[82] C. Schneider , « From LLM to agentic AI: prompt injection got worse » (« Promptware Kill Chain C. Schneider, 2026). https://christian-schneider.net/blog/prompt-injection-agentic-amplification/

[83] OWASP, « LLM01:2025 Prompt Injection », Top 10 for LLM Applications 2025. https://genai.owasp.org/

[84] S. Mishra, S.P. Morgan, Palo Alto Networks Unit 42, « Persistent memory as accelerant for stateful, delayed-execution attacks », février 2026. Réf. The Hacker News.

[85] CybersecurityNews, « OpenClaw Log Poisoning Vulnerability Allows Malicious Content Injection », février 2026. https://cybersecuritynews.com/openclaw-log-poisoning-vulnerability/

[86] HiddenLayer, démonstration injection via HEARTBEAT.md dans OpenClaw, février 2026. Réf. TechInformed.

[87] J. O'Reilly (Dvuln), démonstration manipulation ranking ClawHub, février 2026. Réf. SecurityWeek/Barrack.ai.

[88] Hudson Rock, « Infostealer Steals OpenClaw AI Agent Configuration Files and Gateway Tokens » (soul.md, openclaw.json, device.json), février 2026. Via The Hacker News. https://thehackernews.com/2026/02/infostealers-steal-openclaw-configs.html

[89] Cryptika, « How Infostealers Compromise OpenClaw Agent Identity: Technical Analysis », février 2026.

[90] Barrack.ai, « OpenClaw is a Security Nightmare — Here's the Safe Way to Run It » (absence chiffrement au repos, absence hardware binding, absence audit API), février 2026. https://blog.barrack.ai/openclaw-security-vulnerabilities-2026/

[91] DeepStrike, « 500 000 deepfakes en 2023, 8 millions en 2025, croissance 900 %/an. Augmentation 704 % face-swap » ; Gartner, « 30 % des entreprises ne feront plus confiance aux solutions IDV autonomes d'ici 2026 ». Réf. Keepnet Labs.

[92] Deloitte Center for Financial Services, projection fraude IA : de 12,3 Mds$ (2023) à 40 Mds$ (2027), TCAC 32 %. Réf. Keepnet Labs.

[93] S. Lyu (University at Buffalo), « 2026 will be the year you get fooled by a deepfake ». Voice cloning : « indistinguishable threshold » franchi. Fortune, décembre 2025. https://fortune.com/2025/12/27/2026-deepfakes-outlook-forecast/

[94] Cyble, « Deepfake-as-a-Service Exploded in 2025 » (DaaS), décembre 2025. Pertes fraude US : 12,5 Mds$ en 2025. https://cyble.com/knowledge-hub/deepfake-as-a-service-exploded-in-2025/

[95] Keepnet Labs, « Deepfake Statistics & Trends 2026 ». Cas Arup (février 2024, 25,6 M$), CEO fraud ciblant 400 entreprises/jour. https://keepnetlabs.com/blog/deepfake-statistics-and-trends

[96] World Economic Forum / R. Greig (CIO Arup), « Cybercrime: Lessons learned from a $25m deepfake attack », février 2025. https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/

[97] Cogent Info, « Deepfake Onslaught: Why 2026 Will Demand Enterprise-Grade Defense Now ». https://www.cogentinfo.com/resources/deepfake-onslaught-why-2026-will-demand-enterprise-grade-defense-now

[98] CISA, « KEV Catalog — CVE-2024-21762 Fortinet FortiOS Out-of-Bounds Write » (CVSS 9.8). https://www.cisa.gov/known-exploited-vulnerabilities-catalog

[99] Fortinet PSIRT, « FG-IR-24-015 : FortiOS SSL-VPN Buffer Overflow », février 2024. https://www.fortiguard.com/psirt/FG-IR-24-015

[100] Mandiant / Google Threat Intelligence, « UNC3886: Chinese Espionage Actor Targeting FortiOS Firmware » (implant firmware persistant), 2024. https://www.mandiant.com/resources/blog/unc3886

[101] MITRE ATT&CK, « T1036 Masquerading » et « T1071.001 Web Protocols » (mimicry trafic légitime). https://attack.mitre.org/techniques/T1036/

[102] PurpleSec, « Deepfake Finance Fraud Statistics 2025 » (pertes Q1 2025 : 200 M$ Amérique du Nord). https://purplesec.us/learn/deepfake-statistics/

[103] OWASP, « Top 10 for Agentic Applications 2026 » (Agentic Supply Chain, Uncontrolled Autonomy). https://owasp.org/www-project-top-10-for-agentic-applications/

[104] Adversa.ai, « OpenClaw Security Analysis: AI Agent Vulnerabilities » et « Lethal Trifecta », février 2026. https://adversa.ai/blog/openclaw-security/

[105] Menlo Security, « Predictions for 2026: Why AI Agents Are the New Insider Threat », janvier 2026. https://www.menlosecurity.com/blog/predictions-for-2026

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

[107] MxD / M. Tanji, « Warning: The AI Deepfake Danger Intensifies » (processus haute sécurité, élimination biométrie unique), décembre 2025. https://www.mxdusa.org/news/warning-the-ai-deepfake-danger-intensifies/

[108] Cibersafety, « Deepfakes and social engineering attacks: how digital fraud is evolving in 2025 », décembre 2025. https://cibersafety.com/en/deepfakes-social-engineering-fraud-2025/

[109] MITRE ATLAS, « AML.T0051.001 LLM Prompt Injection — Indirect ». https://atlas.mitre.org/techniques/AML.T0051.001

[110] Aikido.dev, « We audited OpenClaw's agent skills and found critical security issues » (analyse sécurité skills, recommandations sandboxing), février 2026. https://www.aikido.dev/blog/openclaw-audit

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*

[7] Techzine, « Over 40,000 OpenClaw agents vulnerable » (SecurityScorecard), février 2026. https://www.techzine.eu/news/security/138633/
→ *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*

[13] Hudson Rock, « Infostealer Steals OpenClaw AI Agent Configuration Files and Gateway Tokens », via The Hacker News, février 2026.
→ *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*

[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*

[159] Cisco, « State of AI Security 2025 Report » (34 % entreprises avec contrôles IA spécifiques, <40 % tests réguliers).
→ *Defined in Phase 5*

[160] OWASP, Top 10 for Agentic Applications 2026. Sandboxing agentique, moindre privilège, audit trafic sortant.
→ *Defined in Phase 5*