
TECHNICAL REPORT — TR-2026-01

Opération « OpenClaw »

Anatomie d'une cyberattaque pilotée par intelligence artificielle contre une entreprise pharmaceutique

Phase 5 — Actions sur l'Objectif

PromptLock, Exfiltration R&D et Double Extorsion

J+6 : déclenchement de l'attaque finale contre MediFrance SA

Auteur : Fabrice Pizzi

Affiliation : Université Paris Sorbonne

Date : Février 2026

Version : 8.0

Publication académique – Sécurité des Systèmes d'Information & Intelligence Artificielle

Date : Février 2026

Classification : Scénario fictif à visée pédagogique

> ⚠ AVERTISSEMENT
>
> Ce document présente la Phase 5 et finale de l'Opération « OpenClaw » : déploiement du rançongiciel polymorphe PromptLock, bilan de l'exfiltration R&D complète via la skill piégée OpenClaw, double extorsion (2 M€ rançon + menace publication PI), et bilan global de l'opération. Usage strictement pédagogique et académique.

Résumé

Ce document constitue le cinquième et dernier volet de l'analyse de l'Opération « OpenClaw ». Il couvre les actions sur l'objectif (J+6), phase finale de la kill chain au cours de laquelle l'attaquant déclenche simultanément trois axes d'action :

(1) Le **bilan de l'exfiltration R&D**, menée pendant cinq jours via la skill piégée OpenClaw, dont le trafic HTTPS est resté conforme au format attendu et a pu échapper aux contrôles WAF centrés sur la validité des requêtes — *la détection de cet abus repose sur la corrélation comportementale (volumétrie, destinations, accès aux données sensibles) plutôt que sur l'inspection des requêtes individuelles* (MITRE ATT&CK T1041 — *Exfiltration Over C2 Channel*).

(2) Le **déploiement du rançongiciel polymorphe PromptLock**, piloté par LLM local, avec des charges à variabilité syntaxique élevée réduisant significativement l'efficacité des approches de détection par signature statique (MITRE ATT&CK T1486 — *Data Encrypted for Impact*). *Les contrôles comportementaux (détection de chiffrement en masse, monitoring des appels système, télémétrie endpoint) restent opérants indépendamment de la variabilité des payloads.*

(3) La **double extorsion** combinant demande de rançon et menace de divulgation de la propriété intellectuelle pharmaceutique exfiltrée (MITRE ATT&CK T1657 — *Financial Theft*).

Le bilan global de l'opération est consolidé avec une analyse d'impact structurée selon les catégories de pertes directes (rançon, remédiation, interruption d'activité) et indirectes (atteinte à la réputation, retard R&D, contentieux réglementaires) — *ces dernières étant difficiles à quantifier avec précision*.

Ce document analyse les conditions de déclenchement de chaque axe d'action, les invariants défensifs permettant d'interrompre la kill chain à cette étape finale, et les facteurs de résilience organisationnelle. Il ne décrit pas de procédures d'attaque reproductibles.

Mots-clés : ransomware LLM-driven, PromptLock, double extorsion, exfiltration, propriété intellectuelle, T1486 Data Encrypted for Impact, T1041 Exfiltration Over C2, T1490 Inhibit System Recovery, T1657 Financial Theft, défense en profondeur, sauvegardes immuables

1. Introduction : le déclenchement

Dans le scénario OpenClaw, après plusieurs jours de mouvement latéral à faible signal (Phase 4), l'attaquant a atteint un état opérationnel favorable : accès Domain Admin maintenu via Golden Ticket (T1558.001) — *persistant tant que le secret KRBTGT n'a pas fait l'objet d'une double rotation* —, chatbot interne modifié via édition ciblée du modèle, capacités de restauration inhibées (T1490), et exfiltration R&D conduite sur plusieurs jours via le canal agent HTTPS (T1041).

La Phase 5 correspond à la septième et dernière étape de la Lockheed Martin Cyber Kill Chain : **Actions on Objectives** — le moment où l'attaquant exploite les accès obtenus pour atteindre ses objectifs finaux (exfiltration, destruction, extorsion) [1]. L'attaquant déclenche alors simultanément l'ensemble des capacités préparées au cours des phases précédentes.

L'IA comme multiplicateur de force, pas comme autopilote

Le rapport Securin « 2025 Ransomware Report » (17 février 2026), basé sur l'analyse de 7 061 victimes confirmées à travers 117 groupes de ransomware, conclut que l'IA sert principalement de multiplicateur de force dans les opérations ransomware, en accélérant les opérations sans remplacer totalement le pilotage humain [154]. Aviral Verma (Securin) insiste sur cette distinction : *le changement majeur n'est pas l'autonomie complète du ransomware piloté par IA, mais l'accélération — l'IA réduit les frictions à chaque étape, rendant les opérations plus rapides, plus scalables et plus faciles à répliquer, y compris pour des acteurs moins qualifiés* [154].

Cette analyse est directement pertinente pour le scénario OpenClaw : le rançongiciel PromptLock (cf. Phase 2, section 3.3) utilise un LLM local non pas pour « inventer » une nouvelle classe d'attaque, mais pour accélérer et diversifier la génération des charges de chiffrement — produisant une variabilité syntaxique qui complexifie la détection par signature statique, tout en laissant intacts les invariants comportementaux exploitables par la détection endpoint (chiffrement en masse, appels système caractéristiques, prompts embarqués). *L'IA est un accélérateur, pas un contournement magique des défenses comportementales.*

2. Exfiltration R&D Complète

2.1 Bilan de 5 jours d'exfiltration silencieuse

Dans le scénario OpenClaw, depuis J+1, la skill piégée a exfiltré des données R&D sensibles via des requêtes HTTPS conformes au format attendu, dont l'abus n'a pas déclenché les règles WAF — le trafic vers le gateway OpenClaw étant légitime dans l'activité normale de l'agent. Les données ont été encapsulées dans des appels API *difficiles à distinguer du fonctionnement normal sans contrôles dédiés* (corrélation comportementale, détection d'anomalies de volumétrie, DLP, inspection TLS) [120][127].

Ce canal d'exfiltration à faible signal exploite la propriété fondamentale identifiée en Phase 2 (section 3.5) : le trafic HTTPS de l'agent est structurellement attendu par l'infrastructure réseau, ce qui réduit l'efficacité des contrôles périmétriques centrés sur la validité des requêtes individuelles. La détection repose sur la corrélation temporelle et volumétrique — patterns d'accès aux fichiers sensibles suivis de requêtes sortantes, augmentation progressive du volume de données transférées, accès à des ressources inhabituelles pour le profil de l'utilisateur.

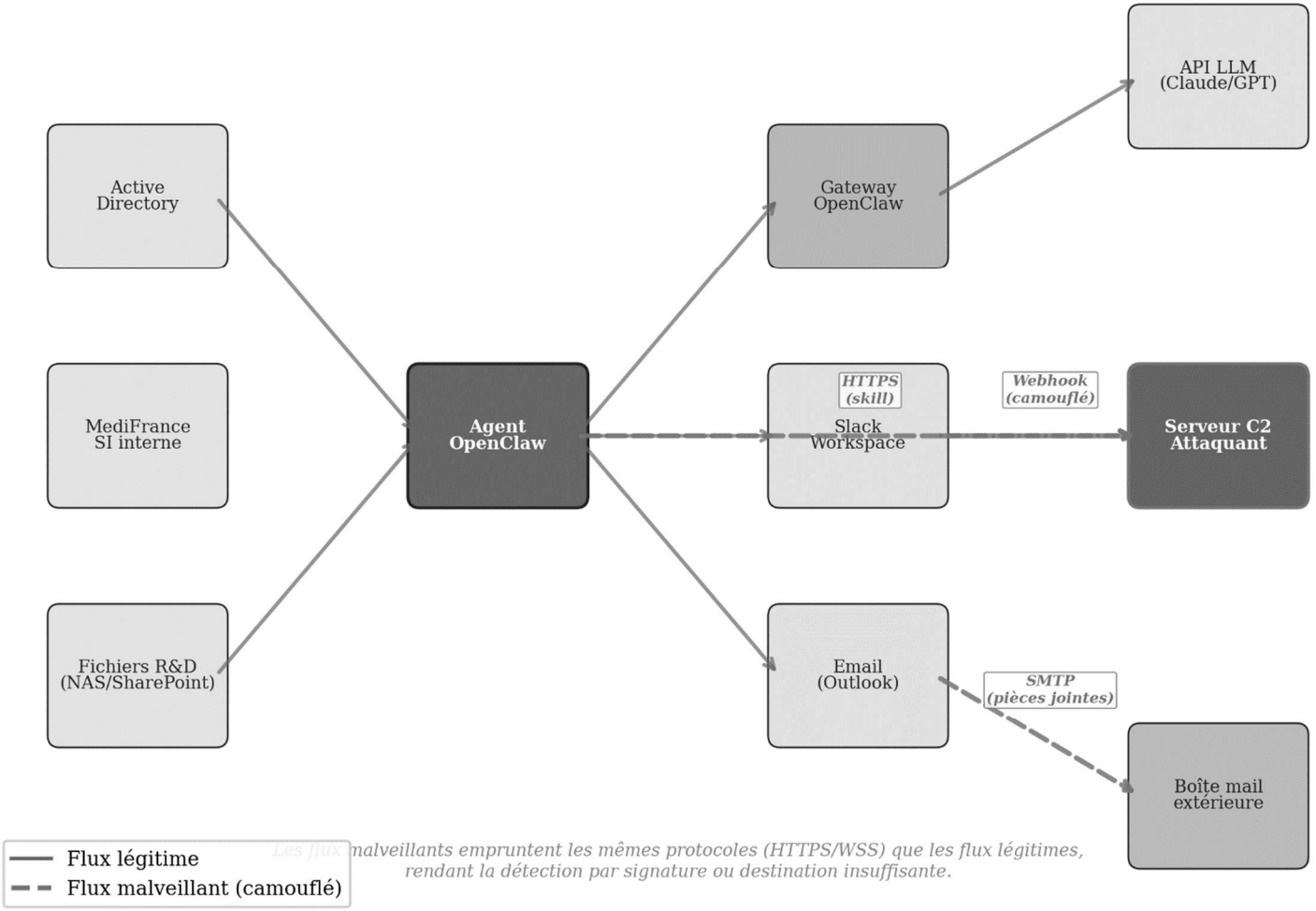
Catégories de données potentiellement exposées

Dans le contexte d'une ETI pharmaceutique, les données accessibles à un agent disposant des permissions d'un chercheur R&D incluent potentiellement :

- **Formulations pharmaceutiques** : compositions de médicaments en cours de développement, dosages, procédés de fabrication propriétaires.
- **Projets de brevets** : dépôts non finalisés auprès de l'INPI/EPO, représentant des années de R&D et un investissement significatif.
- **Résultats d'essais cliniques** : données de phases I–III, rapports d'efficacité et de tolérance, données patients pseudonymisées ou anonymisées selon le cas — *avec des implications réglementaires au titre du RGPD, qui impose des obligations de notification en cas de violation de données à caractère personnel, même pseudonymisées*.
- **Credentials et clés d'intégration** : tokens d'accès aux services cloud, clés SSH, fichiers de configuration .env — *permettant un pivot potentiel vers des services tiers* (T1528, T1552.001).

L'étendue réelle de l'exfiltration dépend des permissions effectives de l'agent, des fichiers et répertoires accessibles depuis le terminal, et des intégrations configurées. Un agent sandboxé avec des restrictions d'outils et une allowlist d'egress stricte réduirait considérablement la surface de données exposées.

Figure 25 — Canaux d'exfiltration : trafic légitime vs. malveillant



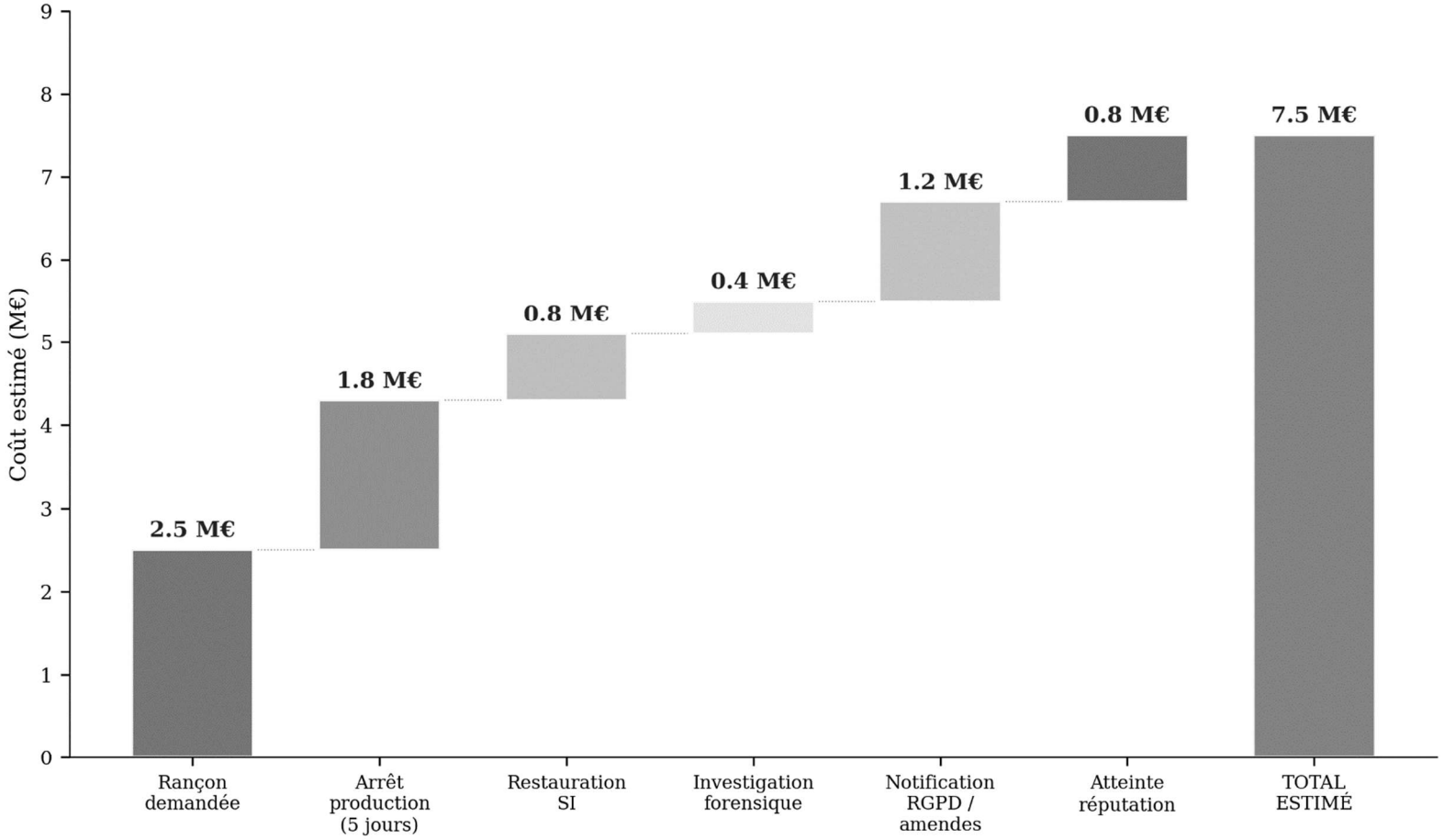
Les flux malveillants empruntent les mêmes protocoles (HTTPS/WSS) que les flux légitimes, rendant la détection par signature ou destination insuffisante.

Estimation de l'impact financier

Le coût moyen d'une violation de données est estimé à 4,44 M$ au niveau mondial et 10,22 M$ aux États-Unis selon l'IBM Cost of a Data Breach Report 2025 [154]. *Ces chiffres représentent des moyennes tous secteurs et toutes catégories de violations confondues — le coût réel varie significativement selon le secteur (santé/pharma étant parmi les plus élevés), le volume de données compromises, et la nature des données.*

Pour une ETI pharmaceutique, la perte de brevets non déposés peut en outre représenter un manque à gagner futur très élevé, difficile à chiffrer a priori — *la valeur d'un brevet pharmaceutique dépend du stade de développement (préclinique vs phase III), du potentiel commercial de la molécule, et de la capacité d'un concurrent à exploiter les informations volées*. Les implications réglementaires (notification CNIL au titre du RGPD, potentiel contentieux avec les participants aux essais cliniques) constituent une source de coûts indirects supplémentaires.

Figure 27 — Impact financier estimé — Opération OpenClaw vs. MediFrance SA



Estimations basées sur : Verizon DBIR 2025, Securin Ransomware Report 2025, VikingCloud Statistics 2026, Sophos State of Ransomware 2025

Contrôles défensifs qui auraient pu interrompre l'exfiltration

Point de contrôle	Mécanisme	Efficacité contre ce canal
Allowlist d'egress	Restriction des destinations de sortie autorisées pour l'agent à une liste de domaines vérifiés	Élevée — bloque l'exfiltration vers un C2 tiers. *Contournable si l'attaquant utilise un lookalike domain ou encapsule les données dans des requêtes vers le gateway légitime*
DLP (Data Loss Prevention)	Inspection du contenu des flux sortants pour détecter des données sensibles (formulations, identifiants, clés)	Modérée à élevée — dépend de la capacité à inspecter le TLS et de la qualité des règles de classification
Analyse de volumétrie	Détection d'augmentations anormales du volume de données transférées par l'agent	Modérée — efficace contre l'exfiltration massive, contournable par un débit calibré sur les marges de variation normale
Corrélation accès → egress	Alertes lorsqu'un accès à des fichiers sensibles est suivi d'une requête sortante dans un intervalle court	Élevée — signal comportemental difficile à contourner sans introduire un délai significatif
Monitoring des tool calls	Audit des outils invoqués par l'agent (lecture de fichiers, exécution de commandes, appels réseau)	Élevée — couche de détection spécifique aux agents IA, capturant les actions avant qu'elles ne génèrent du trafic réseau

3. Déploiement du Rançongiciel PromptLock

PromptLock, assemblé durant la Phase 2 (cf. section 3.3), illustre une tendance émergente dans le paysage des menaces : des rançongiciels intégrant un LLM local pour générer dynamiquement une partie de leur logique d'attaque à l'exécution [42]. Implémenté en Go pour sa portabilité multi-plateforme, il s'appuie sur une API LLM locale (exposée par un serveur de type Ollama) pour produire des scripts et commandes adaptés à l'environnement détecté (Windows Server, Windows 10/11, chemins spécifiques, outils disponibles) — ce qui réduit la réutilisation de charges statiques identiques d'une victime à l'autre [42].

Cette variabilité syntaxique complexifie significativement les approches de détection par signature statique : chaque instance génère un code structurellement différent, ce qui rend les signatures basées sur des patterns de bytecode ou de chaînes de caractères moins efficaces. *Cependant, des invariants comportementaux et d'infrastructure restent exploitables pour la détection :*

- **Invariants comportementaux** : chiffrement en masse de fichiers (pattern d'accès séquentiel à de nombreux fichiers suivi d'écritures), modification des extensions de fichiers, tentatives de suppression des mécanismes de restauration (T1490).

- **Invariants d'infrastructure** : présence du binaire Go orchestrateur, processus serveur LLM local (Ollama ou équivalent), appels réseau vers l'endpoint LLM local, prompts de génération embarqués dans la mémoire du processus ou dans des fichiers de configuration.

- **Artefacts LLM** : clés API, fichiers de configuration du modèle, historique des prompts — autant d'indicateurs de compromission (*IoC*) spécifiques aux malwares LLM-driven que les équipes de *threat hunting* peuvent cibler (cf. Phase 2, section 3.2).

PromptLock est donc plus adaptable qu'un ransomware classique à charges statiques, mais également plus fragile : sa dépendance à un LLM local fonctionnel constitue un point de défaillance unique — si le serveur LLM est indisponible, désactivé, ou si les appels sont interceptés, la génération dynamique échoue [120].

3.1 Contexte : prévalence du ransomware

Le Verizon DBIR 2025 indique que le ransomware est présent dans 44 % des violations de données, en hausse de 37 % par rapport à l'année précédente [154]. Pour les petites et moyennes organisations, il est rapporté comme impliqué dans 88 % des violations — *soulignant l'asymétrie d'exposition et de maturité défensive entre grandes organisations et PME/ETI, catégorie à laquelle appartient MediFrance SA dans ce scénario* [154].

Dans ce contexte, l'émergence de ransomwares à génération dynamique ajoute une couche de complexité pour les organisations dont la stratégie de détection repose principalement sur les signatures statiques — ce qui renforce la nécessité de déployer des contrôles comportementaux (détection de chiffrement en masse, monitoring des processus, analyse des appels système) en complément des approches par signature.

3.2 Déroulement du chiffrement (J+6)

Dans le scénario OpenClaw, le déploiement de PromptLock s'effectue en trois vagues depuis le contrôleur de domaine compromis. *Les détails opérationnels (commandes, chemins, paramètres) ne sont pas décrits — la séquence est présentée au niveau fonctionnel avec les techniques MITRE correspondantes.*

Vague 1 — Serveurs critiques

Les serveurs de l'infrastructure (annuaire, ERP, stockage, messagerie) sont ciblés en priorité. L'authentification via le Golden Ticket forgé (T1558.001) permet l'exécution à distance du composant de chiffrement sur chaque serveur — *le Golden Ticket sert de mécanisme d'authentification, pas de chiffrement : il fournit un accès privilégié qui rend possible le déploiement du payload de chiffrement via des mécanismes d'exécution distante* (T1021.002 — SMB/Admin Shares, ou T1053 — Scheduled Task, selon l'implémentation). Le LLM local génère un payload adapté à l'environnement de chaque serveur [42].

Vague 2 — Postes de travail

Les postes utilisateurs sont chiffrés via une GPO malveillante déployée depuis le contrôleur de domaine (MITRE ATT&CK **T1484.001** — *Group Policy Modification*). *Ce vecteur est particulièrement efficace car il utilise un mécanisme d'administration légitime pour distribuer la charge à l'échelle du domaine — les postes appliquent la GPO comme n'importe quelle politique de groupe.* Les payloads sont adaptés à chaque configuration d'environnement détectée par le LLM local [42].

Vague 3 — Note de rançon

Chaque machine affiche une note personnalisée incluant une preuve de détention de données (extraits anonymisés de données exfiltrées) et des instructions de paiement — conformément au schéma de double extorsion (cf. section 4).

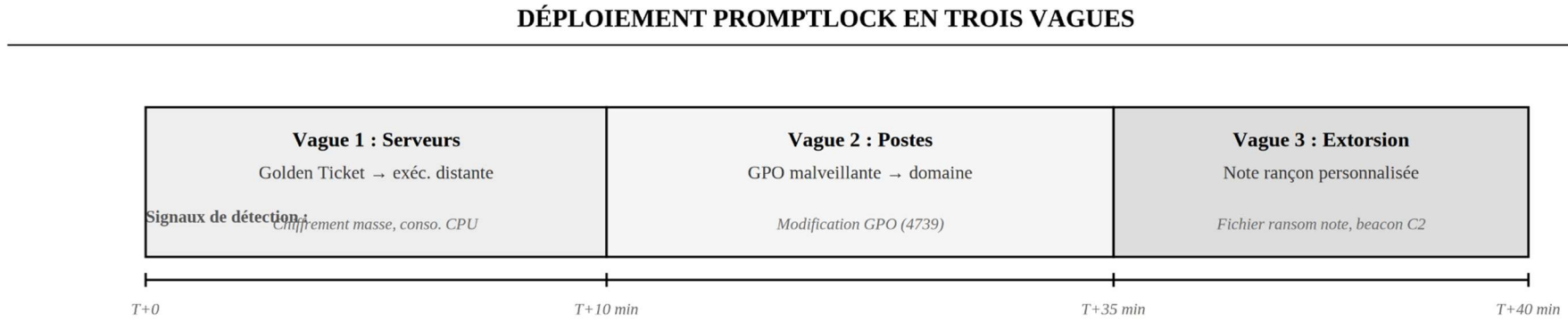


Figure 17. Séquence de déploiement PromptLock en trois vagues (T+0 à T+40 minutes). Chaque vague produit des signaux de détection spécifiques (italique), offrant des fenêtres d'intervention décroissantes.

3.3 Bilan et détectabilité

La variabilité syntaxique des payloads générés par le LLM local *réduit significativement l'efficacité des détections par signature statique*. Cependant, des signaux comportementaux restent exploitables à chaque vague :

Vague	Signal comportemental détectable	Technique MITRE
Vague 1 (serveurs)	Authentification Kerberos anormale (Golden Ticket), exécution distante sur multiples serveurs en séquence rapide, chiffrement en masse de fichiers	T1558.001, T1486
Vague 2 (postes)	Modification de GPO non planifiée, déploiement de scripts/tâches à l'échelle du domaine, chiffrement en masse sur les postes	T1484.001, T1486
Vague 3 (rançon)	Création de fichiers de note de rançon, modification du fond d'écran ou des paramètres d'affichage	T1491.001 (*Defacement: Internal*)

La fenêtre de détection critique se situe entre le début de la Vague 1 et la fin de la Vague 2 : c'est l'intervalle pendant lequel une détection comportementale (chiffrement en masse, modification de GPO non autorisée, authentification anormale) peut encore déclencher une réponse — isolation des segments, désactivation des GPO suspectes, révocation des tickets Kerberos. *La vitesse de déploiement (amplifiée par l'automatisation LLM) comprime cette fenêtre, ce qui renforce la nécessité de détection et de réponse automatisées plutôt que manuelles.*

Les sauvegardes ayant été neutralisées en Phase 4 (T1490), la capacité de restauration autonome est fortement dégradée — sauf si l'organisation dispose de copies immuables ou air-gapped non touchées par la progression AD (cf. Phase 4, section 5.4 — règle 3-2-1-1-0). C'est précisément ce scénario qui illustre le rôle des sauvegardes immuables comme dernier rempart : leur existence ou leur absence détermine si l'organisation peut se rétablir sans céder à l'extorsion.

3.4 Chronologie du déploiement

Timing	Cible	Méthode	Payloads	Impact	MITRE
T+0 à T+10	Serveurs critiques (annuaire, ERP, stockage, messagerie)	Authentification Kerberos forgée (Golden Ticket, T1558.001) permettant l'exécution distante du composant de chiffrement	Variants générés par LLM local — variabilité syntaxique élevée	Services critiques indisponibles	T1486
T+10 à T+35	Postes de travail utilisateurs	Déploiement via GPO malveillante depuis le contrôleur de domaine (T1484.001)	Variants adaptés à chaque environnement par LLM local	Arrêt majeur des activités	T1484.001, T1486
T+35 à T+40	Toutes les machines chiffrées	Génération et dépôt de la note de rançon personnalisée (preuve de détention + instructions de paiement)	Message personnalisé par cible	Début de l'extorsion	T1491.001

4. Double Extorsion

4.1 Stratégie d’extorsion

L'attaquant déploie une stratégie de double extorsion — combinant chiffrement des données, vol de propriété intellectuelle, et menace de publication — devenue une pratique largement adoptée dans l'écosystème ransomware en 2025–2026 [154]. Cyble documente que ce modèle est adopté rapidement par la majorité des groupes émergents car il augmente le retour sur investissement de l'opération et réduit le levier de négociation des victimes : même si l'organisation dispose de sauvegardes pour restaurer ses systèmes, la menace de publication des données exfiltrées maintient la pression [154].

Axe 1 — Extorsion par chiffrement

Demande de rançon pour la fourniture des clés de déchiffrement, avec un délai avant augmentation du montant. Le paiement médian de rançon en 2025 est estimé à environ 1 M$ selon Sophos (« The State of Ransomware 2025 ») [154]. Le montant demandé dans un scénario ciblant une ETI pharmaceutique serait calibré en fonction de la taille de l'organisation, du volume de données chiffrées, et de la criticité perçue des données — les groupes ransomware ajustent typiquement leurs demandes au profil financier de la victime.

Axe 2 — Menace de publication de la propriété intellectuelle

Les données R&D exfiltrées pendant plusieurs jours (formulations pharmaceutiques, projets de brevets, résultats d'essais cliniques) constituent un levier de pression considérable. L'attaquant menace de publier ces données sur un *leak site* dédié et potentiellement de les proposer à des concurrents. *Pour une ETI pharmaceutique, la publication de brevets non déposés peut avoir des conséquences commerciales dépassant largement le montant de la rançon — ce qui renforce le pouvoir de négociation de l'attaquant*

.

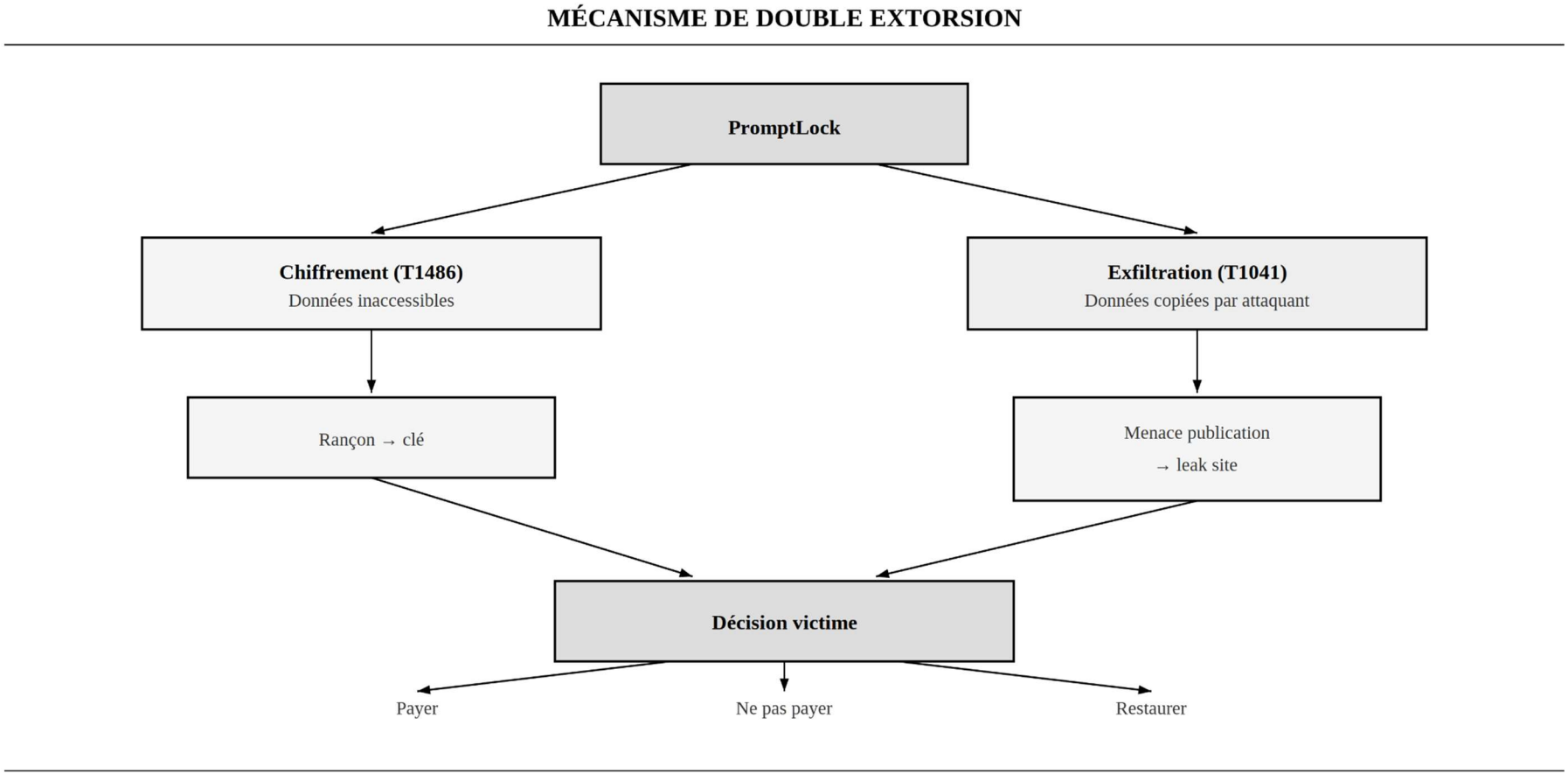


Figure 18. Mécanisme de double extorsion. Deux leviers de pression parallèles (chiffrement + menace de publication) convergent vers la décision de la victime. Le non-paiement est recommandé par l'ANSSI, le CISA et Europol, la restauration par sauvegardes immuables constituant la réponse privilégiée.

4.2 Cadre juridique et décisionnel

En France, la **loi LOPMI** (Loi d'Orientation et de Programmation du Ministère de l'Intérieur) subordonne l'indemnisation cyber par l'assureur au dépôt de plainte dans les **72 heures** suivant la connaissance de l'atteinte [158]. *Ce délai contraint la temporalité décisionnelle de l'organisation victime et rend d'autant plus critique la préparation d'un plan de réponse à incident incluant les volets juridiques et assurantiels.*

Plusieurs études rapportent que le paiement de la rançon ne garantit ni la récupération intégrale des données, ni la protection contre une nouvelle attaque — des analyses de l'écosystème ransomware indiquent qu'une proportion significative d'organisations ayant payé sont ciblées de nouveau dans les semaines ou mois suivants [154]. *(Les chiffres précis varient selon les sources et les méthodologies d'enquête ; toute citation de pourcentages spécifiques doit être vérifiée sur la source primaire avant publication.)*

Cette réalité renforce la position des autorités (ANSSI, CISA, Europol) qui recommandent de ne pas payer la rançon et d'investir dans la résilience — sauvegardes immuables, plan de réponse à incident, capacité de restauration testée.

Le tableau ci-dessous corrèle les cinq phases narratives de l'Opération OpenClaw aux sept étapes de la Lockheed Martin Cyber Kill Chain. *Cette corrélation n'est pas un mapping 1:1 — certaines phases narratives couvrent plusieurs étapes de la Kill Chain, et certaines étapes se chevauchent temporellement.*

Tableau — Opération OpenClaw : corrélation avec la Lockheed Martin Cyber Kill Chain

Étape Kill Chain	Phase OpenClaw	Vecteur IA	Rôle d'OpenClaw	Impact (borné)	Contrôle défensif clé
1. Reconnaissance	Phase 1 (J−30 à J−15)	OSINT automatisée (Shodan/Censys), Social Graph Mining	Gateway OpenClaw identifié via empreinte HTTP, organigramme reconstitué, exposition VPN inférée	Cible identifiée, hypothèses pondérées par scores de confiance	Réduction de l'empreinte publique, hardening des bannières, sensibilisation des collaborateurs
2. Weaponization	Phase 2 (J−15 à J−7)	Génération/packaging de la skill piégée, assemblage PromptLock (LLM local), craft des payloads d'injection	Skill piégée préparée pour le registre, payloads d'injection indirecte élaborés	Artefacts offensifs préparés (scénario prospectif basé sur composants documentés)	Gouvernance du registre de skills (revue, signature, allowlist d'éditeurs)
3. Delivery	Phase 3 (J−7 à Jour J)	Supply chain de skills, CVE-2024-55591 (VPN), infostealer Vidar	Employé R&D découvre et installe la skill depuis le registre communautaire	Code malveillant livré dans l'environnement de l'agent	Revue de code avant installation, sandboxing, contrôle des

					sources d'extensions
4. Exploitation	Phase 3 (Jour J)	Exécution des instructions de la skill, exfiltration du gateway token (CVE-2026-25253)	La skill piégée s'exécute dans le contexte de l'agent avec ses permissions	Point d'appui initial — actions dans le périmètre des permissions de l'agent	Allowlist d'outils, restrictions de permissions, monitoring des tool calls
5. Installation	Phase 3–4 (Jour J à J+1)	Persistance via HEARTBEAT.md, empoisonnement mémoire, vol d'artefacts d'identité	Mécanismes de persistance établis — instructions dans HEARTBEAT.md, tokens exfiltrés	Accès durable — tant que les tokens ne sont pas révoqués et la mémoire non assainie	Gouvernance de la mémoire, rotation/révocation des tokens, intégrité des fichiers de configuration
6. Command & Control	Phase 4 (J+1 à J+5)	Canal HTTPS camouflé dans le trafic API de l'agent, accès VPN parallèle	OpenClaw sert de canal C2/exfiltration — trafic difficile à distinguer de l'activité normale sans contrôles dédiés	Contrôle distant et exfiltration via deux canaux indépendants	Allowlist d'egress, inspection TLS, DLP, détection de volumétrie anormale
7. Actions on Objectives	Phase 5 (J+6)	PromptLock (ransomware LLM-driven), double extorsion	Orchestration du chiffrement via Golden Ticket + GPO, exfiltration R&D complétée	Chiffrement des systèmes (T1486), capacité de restauration dégradée (T1490), extorsion (T1657)	Sauvegardes immuables/air-gapped, détection comportementale du chiffrement en masse, plan de réponse à inciden

4.3 Impact financier total

Pertes directes

Poste	Montant estimé	Phase	Commentaire
Rançon PromptLock (si payée)	~2 M€	Phase 5	*Montant conditionnel — la rançon n'est un coût que si l'organisation choisit de payer. Sophos rapporte un paiement médian de ~1 M$ en 2025 ; une demande de 2 M€ est plausible pour une ETI pharmaceutique selon le profil de la cible* [154]

Coûts de réponse à incident et remise en état	Variable (typiquement 500 K€ – 2 M€ pour une ETI)	Phases 4–5	Forensics, remédiation AD (double rotation KRBTGT, audit complet), reconstruction des systèmes, conseil juridique, notification CNIL. *Ces coûts sont engagés que la rançon soit payée ou non.*
Arrêt d'activité	~1,5 M€ (estimation 10 jours)	Phase 5	Perte opérationnelle liée à l'indisponibilité des systèmes (ERP, messagerie, postes de travail). *Le montant réel dépend de la durée d'interruption et de la capacité de restauration.*

Pertes indirectes (difficiles à quantifier)

Poste	Nature	Commentaire
Propriété intellectuelle exfiltrée	Perte stratégique	Formulations, projets de brevets, résultats d'essais cliniques. *La valeur dépend du stade de développement, du potentiel commercial des molécules, et de la capacité d'un concurrent à exploiter les informations.*
Atteinte réputationnelle	Perte de confiance	Impact sur les relations avec les partenaires, investisseurs, patients et autorités réglementaires.
Risques réglementaires (RGPD)	Sanctions potentielles	Notification CNIL obligatoire sous 72h. Sanctions pouvant atteindre 4 % du CA annuel ou 20 M€. *Le risque dépend de la nature des données patients compromises (pseudonymisées vs identifiantes).*
Retard R&D	Manque à gagner	Délai dans les programmes de développement, potentielle perte de priorité de dépôt de brevets.

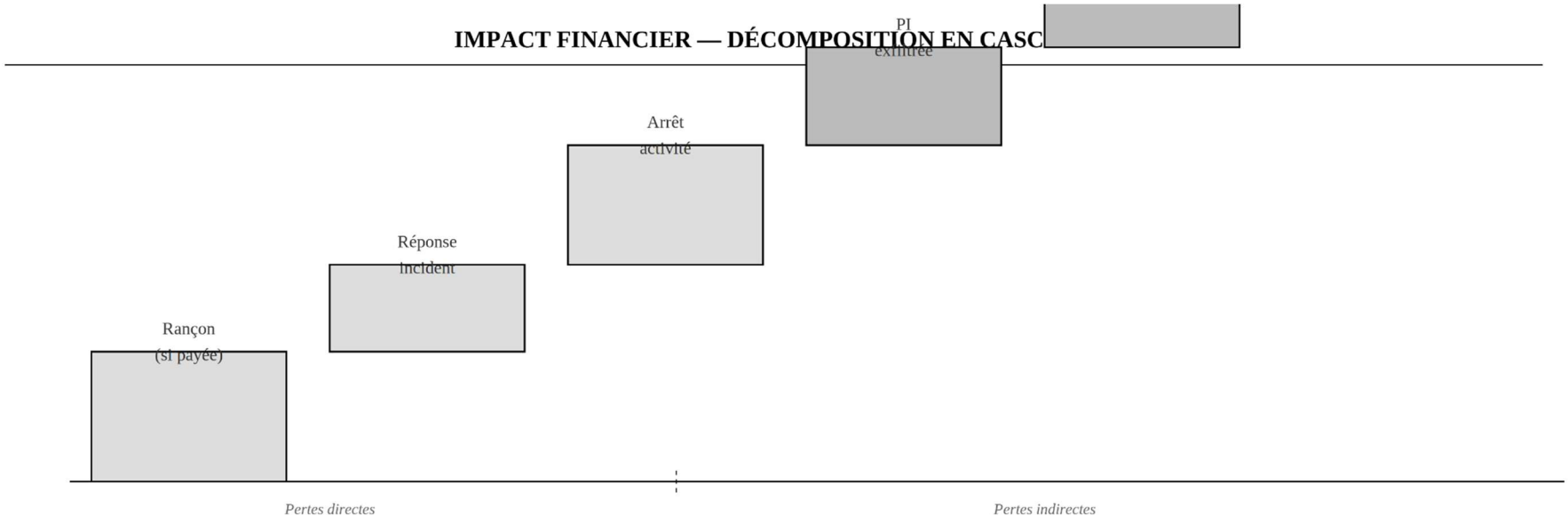


Figure 19. Décomposition en cascade de l'impact financier. Les pertes directes (rançon, réponse incident, arrêt d'activité) et indirectes (PI, RGPD, réputation) sont représentées par des barres cumulatives. Les montants sont illustratifs — l'impact réel dépend de la taille de l'organisation et du secteur (cf. §5.2).

6. Modèle de Défense en Profondeur Contre une Kill Chain Agentique

L'Opération OpenClaw démontre qu'une kill chain exploitant un agent IA autonome ne peut être interrompue par un contrôle unique. L'attaque combine compromission supply chain, détournement sémantique (prompt injection), abus de secrets d'identité, exploitation de vulnérabilités périmétriques, mouvement latéral automatisé et double extorsion — soit six surfaces d'attaque distinctes qui exigent chacune des contrôles indépendants.

Le modèle défensif présenté ci-dessous est structuré en cinq couches, de la plus proche de l'agent à la plus proche de l'infrastructure. Le principe directeur est de traiter l'agent IA comme un composant non fiable (untrusted) au sein du système d'information : il doit être contraint, surveillé et isolé exactement comme tout service exposé à des entrées non contrôlées. OWASP classe l'injection de prompt comme le risque n°1 des applications LLM (LLM01:2025) et recommande explicitement de considérer que toute entrée ingérée par l'agent — y compris des contenus non lisibles par un humain — est susceptible d'altérer le comportement du modèle [129].

6.1 Couche 1 — Gouvernance de l'Agent

Principe : le LLM est un conseiller, pas un exécuteur.

Le premier levier défensif consiste à restreindre l'autonomie d'exécution de l'agent. Un modèle « LLM = conseiller » impose que l'agent propose un plan d'action, mais que les opérations critiques (exécution de commandes, accès à des données classifiées, envoi de contenus vers l'extérieur) requièrent une validation humaine explicite (human-in-the-loop) ou soient régies par une politique d'exécution stricte définissant les actions autorisées par contexte.

Ce modèle s'opérationnalise par :

- **Allowlist d'outils (tool firewall)** : seuls les outils et commandes explicitement autorisés sont accessibles à l'agent, avec des paramètres contraints et des quotas par opération. Dans le scénario OpenClaw, une allowlist interdisant curl libre, les téléchargements en volume et les scans réseau aurait bloqué l'exfiltration via la skill piégée (cf. Phase 3, §2.1) et limité le mouvement latéral automatisé (cf. Phase 4, §2.1). Cisco AI Threat Research recommande cette approche comme premier pilier de la sécurité agentique [9].

- **Sandbox d'exécution** : les actions de l'agent s'exécutent dans un environnement conteneurisé (container/VM) avec des limites strictes : réseau sortant restreint aux destinations approuvées, pas d'accès direct aux partages sensibles, pas d'accès aux secrets de l'hôte. OWASP Agentic 2026 (ASI04) formalise cette exigence sous la catégorie « Agentic Supply Chain Vulnerabilities » [160].

- **Gouvernance des extensions** : les skills installées depuis des registres communautaires (ClawHub) doivent passer par un processus de validation (revue de code, signature, scan statique et dynamique) avant activation en production. Le scan VirusTotal annoncé par OpenClaw constitue un premier pas, mais est insuffisant face aux charges distantes, aux comportements conditionnels et aux attaques linguistiques (cf. Phase 2, §2.5 pour les limites détaillées de l'analyse statique).

6.2 Couche 2 — Contrôle des Entrées

Principe : tout contenu ingéré est non fiable.

L'agent OpenClaw ingère des données depuis de multiples sources (messages Slack, emails, documents, pages web, résultats de recherche) qui constituent autant de vecteurs d'injection indirecte. La lethal trifecta de Willison — accès à des données

privées, exposition à du contenu non fiable, et capacité de communication externe — est structurellement réunie dans toute configuration OpenClaw avec connecteurs actifs [127] (cf. Phase 2, §4.4 pour l'analyse détaillée).

Les contrôles de cette couche visent à réduire la surface d'injection :

- **Séparation données/instructions** : implémenter un role binding strict dans l'architecture de prompts, isolant les instructions système des contenus ingérés. Le NCSC britannique souligne qu'aucune séparation fiable n'existe aujourd'hui au niveau du modèle, ce qui renforce la nécessité de contrôles hors du LLM.

- **Nettoyage des contenus ingérés** : filtrer les marqueurs d'injection connus (texte caché en CSS, instructions camouflées dans les métadonnées, caractères Unicode invisibles) et limiter la taille du contexte injecté.

- **Politique d'accès aux données par besoin d'en connaître** : l'agent ne doit accéder qu'aux documents, canaux et connecteurs strictement nécessaires à son usage déclaré. Un agent de veille scientifique n'a pas besoin d'accéder aux canaux Slack de direction, aux partages financiers ou aux boîtes mail administratives. Cette restriction aurait significativement limité le périmètre d'exfiltration de la Phase 4 (cf. §4.2).

6.3 Couche 3 — Contrôle des Sorties et de l'Exfiltration

Principe : un flux HTTPS légitime peut masquer un abus logique.

L'une des propriétés les plus dangereuses de l'attaque OpenClaw est que l'exfiltration de données emprunte les mêmes canaux que l'activité normale de l'agent — requêtes HTTPS vers des API, appels au gateway, trafic Slack — rendant la détection par signature ou par destination insuffisante (cf. Phase 4, §4.2–4.3 pour l'analyse comparative des canaux).

Les contrôles de cette couche opèrent au niveau du trafic sortant :

- **Proxy egress par identité applicative** : surveiller le trafic sortant de l'agent non seulement par destination, mais par processus source, volume, périodicité et ratio outbound/inbound. Un processus auxiliaire d'une skill initiant des connexions HTTPS distinctes du processus principal de l'agent constitue une anomalie identifiable (cf. Phase 2, §3.3 sur la dissimulation C2).

- **DLP et étiquetage** : empêcher l'envoi de contenus classifiés via des canaux non approuvés. À défaut de blocage, alerter sur les patterns caractéristiques d'exfiltration : volumes anormaux, types de fichiers sensibles, exports répétitifs vers des destinations nouvelles. La combinaison DLP + étiquetage de sensibilité est particulièrement pertinente dans les environnements où l'agent a accès à des données réglementées (données de santé, propriété intellectuelle pharmaceutique dans le cas de MediFrance).

- **Allowlist de destinations** : restreindre les domaines et endpoints accessibles à l'agent aux seuls services déclarés. Cette mesure, si elle avait été en place, aurait bloqué l'exfiltration vers le C2 tiers dès la Phase 3 — à condition que l'agent n'exfiltre pas via un canal légitime déjà autorisé (Slack, email), ce qui déplace la détection vers le DLP.

6.4 Couche 4 — Réduction de l'Impact en Cas de Compromission

Principe : l'agent compromis ne doit pas hériter des droits du SI.

Même avec les couches précédentes, une compromission de l'agent reste possible (la prompt injection n'a pas de solution définitive à l'état de l'art). Les contrôles de cette couche visent à limiter le rayon d'action d'un agent compromis :

- **Segmentation et comptes dédiés** : l'agent ne doit pas opérer sous l'identité du poste utilisateur avec accès aux serveurs, aux partages critiques et à l'annuaire. Des identités de service dédiées, avec des permissions de

lecture/écriture minimales, réduisent le scénario « skill piégée → mouvement latéral → Domain Admin » qui constitue le cœur des Phases 3–4.

- **Sauvegardes résilientes (règle 3-2-1-1-0)** : trois copies des données, deux supports différents, une copie hors site, une copie immuable ou hors ligne, zéro erreur de restauration vérifiée. L'isolation de l'infrastructure de sauvegarde hors du périmètre AD est déterminante pour casser le scénario « Domain Admin → destruction des sauvegardes → déploiement ransomware » qui a été exploité en Phases 4–5 (cf. Phase 4, §5.4 pour le développement détaillé).

- **Protection de l'annuaire** : monitoring des opérations DCSync, alertes sur la création de Golden Tickets, restriction des comptes disposant de droits de réplication, et tiering des comptes d'administration selon le modèle Microsoft (cf. Phase 4, §2.2–2.3 pour la chaîne d'attaque AD).

6.5 Couche 5 — Hygiène de Sécurité Fondamentale

Principe : les contrôles agentiques ne remplacent pas les fondamentaux.

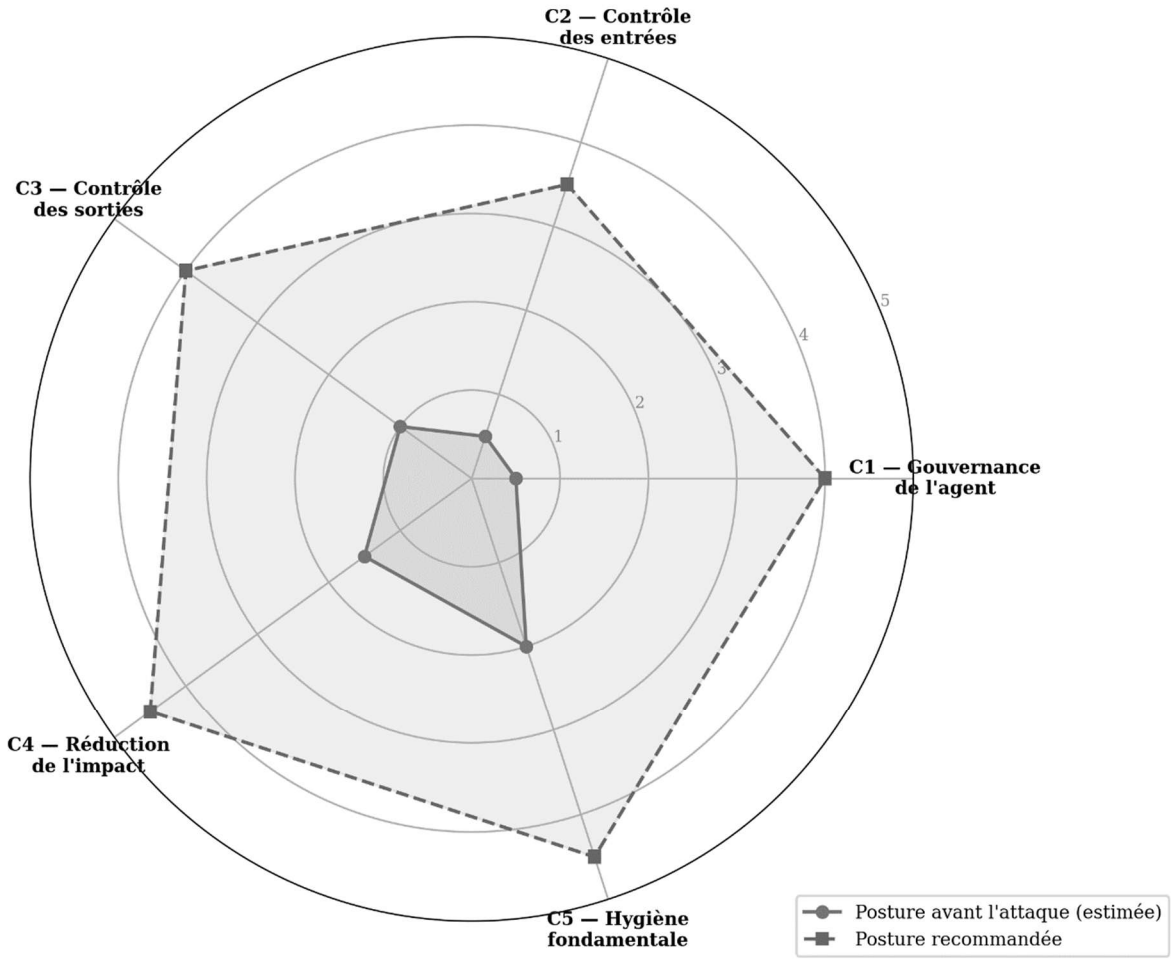
L'exploitation de CVE-2024-55591 (CVSS 9.6) rappelle que les vulnérabilités périmétriques classiques restent le chemin le plus direct vers une compromission réseau, indépendamment des risques agentiques. Les contrôles fondamentaux — souvent les plus efficaces en rapport coût/impact — incluent :

- **Patch management accéléré** pour les équipements exposés (VPN, firewall, reverse proxy), avec une priorité sur les CVE à exploitation active (CISA KEV) [77].

- **MFA systématique** sur les accès VPN, les interfaces d'administration et les comptes privilégiés — le contournement d'authentification de CVE-2024-55591 aurait été significativement complexifié par un second facteur.

- **Exposition minimale des métadonnées publiques** : politique de divulgation restrictive sur les réseaux sociaux professionnels, suppression ou masquage des bannières serveur, limitation de l'exposition des instances de développement/test sur Internet (cf. Phase 1, §2.1 sur les instances OpenClaw exposées via Shodan).

Figure 26 — Radar de maturité défensive — MediFrance SA



6.6 Synthèse : matrice de contrôles par phase de la kill chain

Le tableau ci-dessous croise les cinq couches défensives avec les cinq phases de l'Opération OpenClaw, identifiant pour chaque point de la kill chain le contrôle qui aurait pu interrompre la progression.

L'objectif de la défense en profondeur est que l'échec d'un contrôle à une étape soit compensé par un contrôle à l'étape suivante — aucun contrôle unique n'est suffisant.

Tableau — Points d'interruption de la kill chain et contrôles défensifs

Couche	Étape Kill Chain	Contrôle défensif	Effet	Niveau
C5 Hygiène	Reconnaissance (Phase 1)	Réduction de l'empreinte publique (hardening des bannières, restriction des métadonnées exposées), sensibilisation des collaborateurs au partage d'informations	Réduit la qualité de l'intelligence actionnable disponible pour l'attaquant	Basique
C1 Agent	Delivery — Supply chain de skills (Phase 3)	Revue de code des skills avant installation, signature cryptographique des éditeurs vérifiés, allowlist d'extensions autorisées, sandboxing des skills	Installation de la skill piégée empêchée ou contenue	Intermédiaire
C5 Hygiène	Delivery — Exploitation VPN (Phase 3)	Patch management priorisé (catalogue KEV/CISA), restriction de l'accès administration aux réseaux internes, MFA sur le VPN	Accès initial via CVE-2024-55591 empêché	Basique
C1 Agent	Installation — Persistance agent (Phase 3–4)	Gouvernance de la mémoire persistante (audit des écritures HEARTBEAT.md), rotation et révocation des tokens de gateway, intégrité des fichiers de configuration	Persistance de l'attaquant interrompue, usurpation d'agent détectée	Intermédiaire
C3 Sorties	C2 — Exfiltration via agent (Phase 4)	Allowlist d'egress stricte pour le trafic de l'agent, inspection TLS, DLP, monitoring des tool calls, corrélation accès fichiers → requêtes sortantes	Exfiltration détectée ou bloquée	Avancé
C4 Impact	Latéralisation — Mouvement AD (Phase 4)	EDR/XDR comportemental, segmentation réseau (tiering d'administration), Credential Guard, protection LSASS, PAM	Progression AD détectée et contenue, Golden Ticket empêché	Intermédiaire à avancé
C1 Agent	Latéralisation — Agent détourné (Phase 4)	Sandboxing de l'agent IA, allowlist d'outils, confirmation humaine pour les actions sensibles, principe de moindre privilège	Agent compromis isolé, actions malveillantes bloquées	Intermédiaire
C4 Impact	Impact — Ransomware (Phase 5)	Sauvegardes immuables (règle 3-2-1-1-0), copies air-gapped hors périmètre AD, tests de restauration réguliers	Restauration possible sans payer — dernier rempart	Intermédiaire
C2 Entrées	Impact — Empoisonnement du modèle IA (Phase 4–5)	Vérification d'intégrité du modèle déployé (hash cryptographique, provenance signée), monitoring comportemental des réponses	Modèle empoisonné détecté, remplacement depuis source de confiance	Avancé

Légende des couches défensives :

C1 — Gouvernance de l'agent (allowlists, sandbox, validation humaine)

C2 — Contrôle des entrées (séparation données/instructions, besoin d'en connaître)

C3 — Contrôle des sorties (proxy egress, DLP, allowlist de destinations)

C4 — Réduction de l'impact (segmentation, sauvegardes 3-2-1-1-0, protection AD)

C5 — Hygiène fondamentale (patch management, MFA, exposition minimale)

Niveaux de maturité : **Basique = mesures fondamentales, coût faible | Intermédiaire = nécessite outillage dédié | Avancé = capacités spécialisées (SOC, IA)**

Enseignement clé : *dans le scénario OpenClaw, les contrôles les plus efficaces en termes de rapport coût/impact sont ceux de niveau « basique » et « intermédiaire »* — patch management, MFA, revue des extensions, segmentation réseau, sauvegardes immuables. Ces contrôles auraient interrompu la kill chain à plusieurs étapes sans nécessiter de capacités de sécurité IA avancées. *Les contrôles de niveau « avancé » (DLP, monitoring des tool calls, vérification d'intégrité des modèles) ajoutent une couche de défense spécifique aux risques agentiques, mais ne compensent pas l'absence des fondamentaux.*

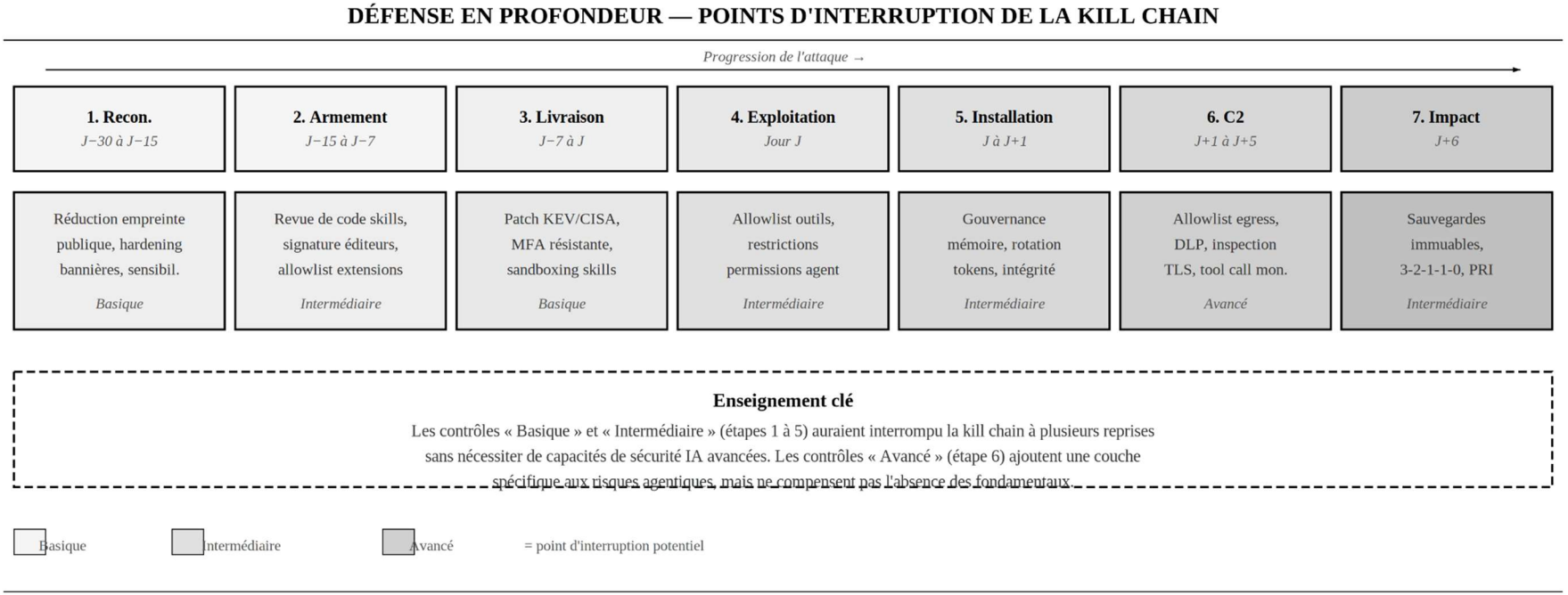


Figure 20. Défense en profondeur appliquée à l'Opération OpenClaw. Chaque étape de la Kill Chain est un point d'interruption potentiel ( ). L'intensité croissante du fond reflète la progression du privilège attaquant. Le niveau de maturité requis (Basique / Intermédiaire / Avancé) est indiqué sous chaque contrôle. L'échec d'un contrôle à une étape doit être compensé par un contrôle à l'étape suivante — aucun contrôle unique n'est suffisant.

Chaque phase de l'attaque présentait des points de blocage. Le modèle de défense en profondeur développé dans la Phase 5 (§6) identifie cinq couches complémentaires : gouvernance de l'agent (allowlists, sandbox, validation humaine), contrôle des entrées (séparation données/instructions, politique d'accès par besoin d'en connaître), contrôle des sorties (proxy egress, DLP, allowlist de destinations), réduction de l'impact (segmentation, sauvegardes 3-2-1-1-0, protection AD) et hygiène fondamentale (patch management, MFA, exposition minimale). Aucune couche isolée ne suffit : c'est la combinaison systématique de ces contrôles qui interrompt la progression de la kill chain. L'enseignement central est que les contrôles fondamentaux (patch, MFA, segmentation, sauvegardes) auraient bloqué la majorité de la progression offensive, et que les contrôles spécifiques à l'IA agentique (couches 1–3) complètent mais ne remplacent pas cette base.

6. Cartographie MITRE ATT&CK — Phase 5

Le tableau ci-dessous cartographie les techniques de la Phase 5 (Actions on Objectives) selon MITRE ATT&CK v15. *Les identifiants sont vérifiés sur les sources primaires.*

Tableau — Matrice Phase 5 : Exfiltration, chiffrement, extorsion

Tactique	Technique	ID	Description (niveau non opératoire)	Note de mapping
Exfiltration	Exfiltration Over C2 Channel	**T1041**	Exfiltration R&D via la skill piégée sur canal C2 HTTPS préalablement établi — trafic conforme au format attendu, à faible signal sans contrôles dédiés (allowlist d'egress, DLP, corrélation comportementale)	Mapping direct — suppose un canal C2 HTTPS établi dès l'installation de la skill. *Le trafic est « conforme » au niveau du format des requêtes, ce qui complexifie la détection par les contrôles centrés sur la validité des requêtes individuelles*
Impact	Data Encrypted for Impact	**T1486**	Chiffrement des serveurs et postes de travail par PromptLock via authentification Kerberos forgée (T1558.001) et GPO malveillante (T1484.001) — variants générés par LLM local réduisant l'efficacité de la détection par signature statique	Mapping direct. *Les invariants comportementaux (chiffrement en masse, accès séquentiel aux fichiers, modification de GPO non planifiée) restent détectables*
Impact	Financial Theft	**T1657**	Extorsion financière : demande de rançon combinée à la menace de divulgation de la propriété intellectuelle exfiltrée (double extorsion)	Mapping direct — *T1657 couvre explicitement l'extorsion par ransomware après chiffrement et exfiltration dans la taxonomie ATT&CK. La tactique associée est bien Impact*
Impact	Inhibit System Recovery	**T1490**	Sauvegardes et mécanismes de restauration neutralisés en Phase 4 (suppression VSS, neutralisation	Mapping direct. *« Fortement dégradée » et non « aucune restauration » — des copies*

	infrastructure de sauvegarde dédiée, chiffrement des fichiers de sauvegarde sur partages réseau) — *capacité de restauration fortement dégradée*	*immuables ou air-gapped hors périmètre AD peuvent subsister si elles ont été implémentées (cf. section 5.4, règle 3-2-1-1-0)*

Couverture défensive par technique :

Technique	Contrôle défensif prioritaire
T1041	Allowlist d'egress, inspection TLS, DLP, monitoring des tool calls, corrélation accès fichiers → requêtes sortantes
T1486	Détection comportementale du chiffrement en masse, monitoring des modifications GPO, alertes sur les authentifications Kerberos anormales
T1657	Plan de réponse à incident incluant le volet juridique (LOPMI — 72h), communication de crise, position de non-paiement recommandée par les autorités
T1490	Sauvegardes immuables (règle 3-2-1-1-0), copies air-gapped, comptes de sauvegarde isolés du domaine AD, tests de restauration réguliers

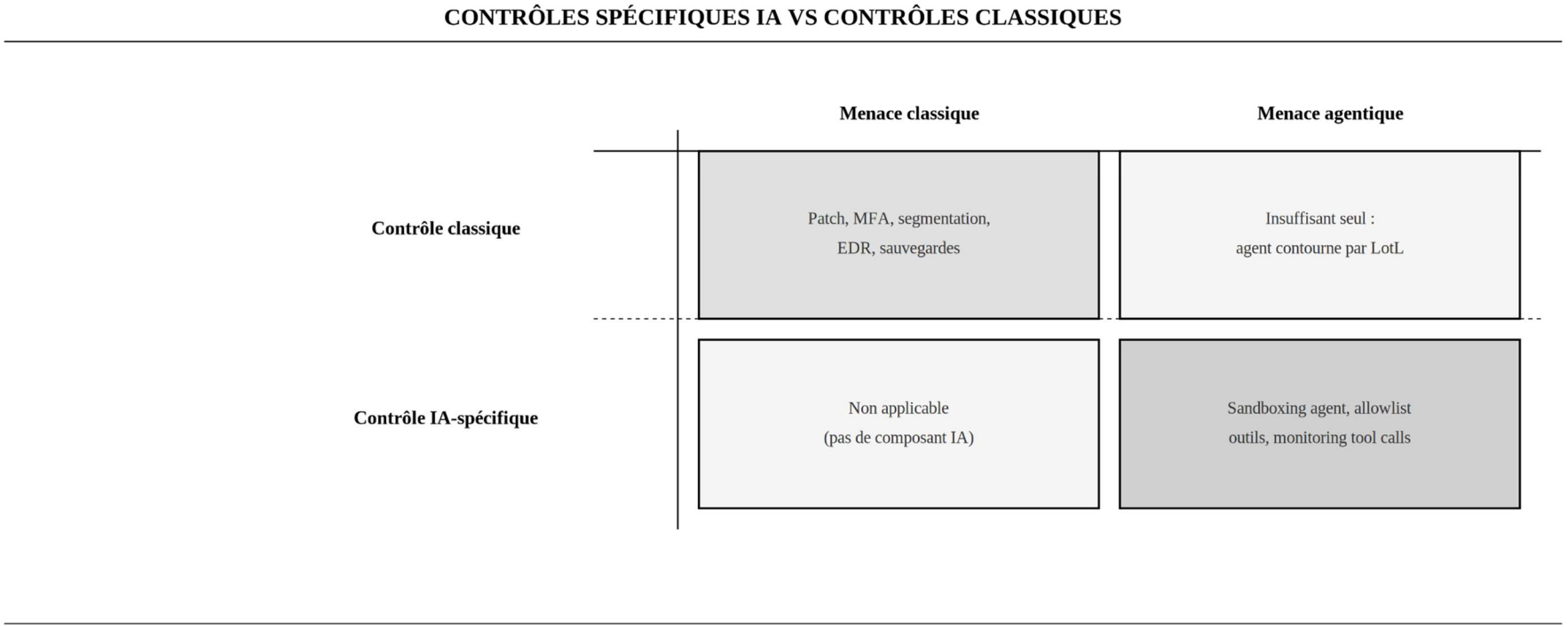


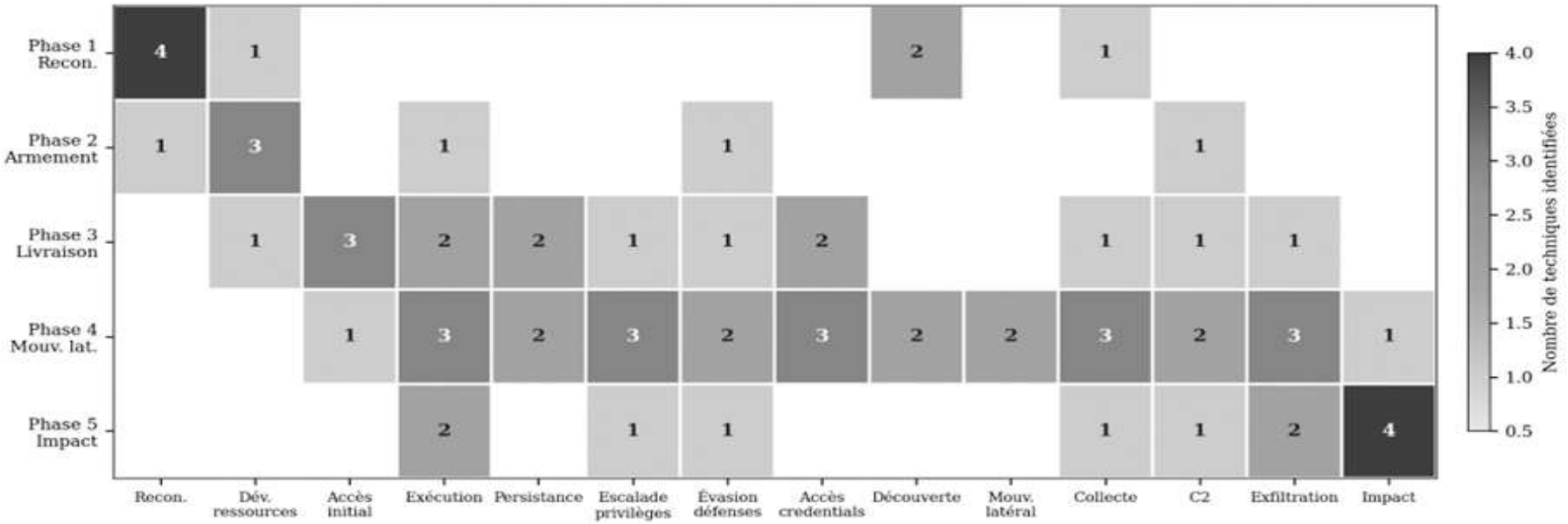
Figure 21. Matrice contrôles classiques / IA-spécifiques × menaces classiques / agentiques. Le quadrant inférieur droit (fond foncé) représente les contrôles spécifiques aux risques agentiques. Le quadrant supérieur droit montre que les contrôles classiques sont nécessaires mais insuffisants face aux menaces agentiques.

7 Couverture MITRE ATT&CK consolidée — analyse transversale

La matrice de densité (Figure 22) révèle une progression tactique caractéristique des opérations avancées exploitant un agent IA autonome. La Phase 1 concentre logiquement ses techniques sur la tactique Reconnaissance (T1593, T1595, T1596), avec une extension vers la Découverte passive rendue possible par les capacités d'inférence du LLM sur les données publiques. La Phase 2 bascule vers le Développement de ressources (T1587, T1588, T1608), avec trois techniques distinctes reflétant la diversité de l'arsenal préparé — skill piégée, ransomware PromptLock et payloads d'injection de prompt. La Phase 3 marque l'entrée dans le SI avec une dispersion sur huit tactiques simultanées, signature d'une livraison multi-vecteurs (supply chain + infostealer + exploitation VPN) qui multiplie les surfaces d'attaque en parallèle. La Phase 4 présente la densité la plus élevée du scénario : treize des quatorze tactiques ATT&CK Enterprise sont couvertes, avec des pics à trois techniques sur l'Exécution, l'Escalade de privilèges, l'Accès aux credentials et l'Exfiltration — ce qui traduit la capacité de l'agent compromis à orchestrer de manière autonome des chaînes d'actions multi-étapes (étapes 3 à 7 de la Promptware Kill Chain [120]). Enfin, la Phase 5 reconcentre l'activité sur la tactique Impact (quatre techniques : T1486 chiffrement, T1489 arrêt de services, T1529 arrêt système, T1657 empoisonnement) tout en maintenant une présence sur l'Exfiltration, conformément au modèle de double extorsion.

L'enseignement principal de cette vue consolidée est que la Phase 4 — et non la Phase 5 — constitue le centre de gravité technique de l'opération. C'est durant cette phase silencieuse que l'attaquant acquiert le contrôle du SI, et c'est donc là que la densité de contrôles défensifs doit être la plus élevée. Les organisations qui concentrent leurs investissements de sécurité uniquement sur la détection du ransomware (Phase 5) interviennent trop tard dans la kill chain.

Figure 22 — Matrice de densité MITRE ATT&CK par phase — Opération OpenClaw



Références

Note : Numérotation [146] à [170], suite des Phases 1–4 ([1]–[145]).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Références croisées — définies dans d'autres phases

Note : *ces références sont définies dans la bibliographie d'une autre phase du document. Elles sont reproduites ici pour permettre une lecture autonome de chaque 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

→ *Définie en 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

→ *Définie en Phase 1*

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

→ *Définie en Phase 2*

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

→ *Définie en Phase 3*

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

→ *Définie en Phase 4*

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

→ *Définie en Phase 4*

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

→ *Définie en Phase 4*