COURS S1-ISI5 – Sécurité des Systèmes d'Information

Intelligence Artificielle et Cybersécurité

Anatomie d'une attaque augmentée – L'IA comme arme et comme bouclier

Enseignant : Fabrice Pizzi
UNIVERSITE SORBONE
MASTER 2 SI-ISI5
Année universitaire 2025-2026

1. Introduction – L'IA, quatrième révolution

L'intelligence artificielle constitue la quatrième révolution technologique et redistribue profondément les cartes dans le domaine de la cybersécurité. Si l'IA générative (GenAI) offre des opportunités majeures pour les défenseurs, elle représente également un multiplicateur de force considérable pour les attaquants.

Ce document explore cette dualité : comment l'IA transforme chaque étape de la Kill Chain du côté offensif, et comment elle renforce les capacités de détection et de réponse du côté défensif. L'écosystème cybercriminel est désormais marqué par une professionnalisation et une spécialisation croissantes, où les outils d'attaque sont créés plus rapidement que les défenses ne sont mises à jour.

L'ANSSI et l'ENISA confirment cette tendance : l'IA accélère à la fois la sophistication des attaques et la capacité de réponse des défenseurs. Nous assistons à une véritable course aux armements numérique.

2. L'IA Offensive – L'arme du cyberattaquant

L'IA est devenue un outil puissant pour les attaquants, permettant d'automatiser, d'améliorer et d'accélérer chaque phase de la cyberattaque. Le modèle RaaS (Ransomware-as-a-Service) combiné à l'IA abaisse considérablement la barrière technique d'entrée.

2.1 Reconnaissance augmentée par IA

L'IA transforme la phase de reconnaissance en permettant une analyse sémantique automatisée de volumes massifs de données publiques pour identifier les cibles à haute valeur et leurs vulnérabilités.

- **Analyse sémantique OSINT :** L'IA analyse des données publiques (LinkedIn, rapports d'entreprise, commits GitHub) pour identifier les cibles à haute valeur et leurs vulnérabilités sociales ou techniques.
- **Cartographie automatisée :** Découverte automatique de surfaces d'attaque via des outils comme Shodan, enrichie par l'IA pour identifier les services cloud mal configurés et les API exposées.
- **Profilage des cibles :** Génération automatique de profils détaillés des employés clés pour préparer des attaques d'ingénierie sociale ultra-ciblées.

2.2 Armement par IA générative (GenAI)

L'IA générative révolutionne la création d'armes cyber en permettant la génération automatisée de code malveillant polymorphe, adapté dynamiquement à chaque cible.

- **Génération de malware polymorphe :** Des LLMs jailbreakés créent des malwares polymorphes ou des scripts d'exploitation multi-plateformes (Bash, PowerShell) contournant les signatures EDR. L'IA exploite des outils déjà présents sur la machine (Living-off-the-Land).
- **Code multi-plateformes :** Production rapide de code malveillant ciblant Windows, Linux et macOS simultanément, rendant la détection par signature traditionnelle obsolète.

- **Abaissement de la barrière d'entrée :** Même des affiliés RaaS avec des compétences techniques limitées peuvent désormais générer des outils offensifs sophistiqués.

2.3 L'ère du spear-phishing hyperpersonnalisé

Le phishing reste le vecteur dominant (60% des cas selon l'ENISA), mais l'IA le rend incomparablement plus dangereux en supprimant les marqueurs traditionnels de détection.

- **GenAI pour le texte :** Création d'e-mails de spear-phishing au style parfait, sans fautes, contextualisés à la cible. Les messages sont indiscernables de communications légitimes.
- **Deepfakes audio et vidéo :** Usurpation de l'identité d'un dirigeant par appel vidéo ou vocal (vishing). Un CEO deepfaké peut demander un virement urgent par visioconférence.
- **Prompt Injection indirecte :** Une charge malveillante est cachée dans une page web que le LLM d'entreprise de la victime va consulter, pouvant mener à une fuite de données.

2.4 Exploitation automatisée – La chasse aux zero-days

L'IA compresse radicalement le cycle de vie des vulnérabilités zero-day, passant de plusieurs mois à quelques heures entre la découverte et l'exploitation.

- **Fuzzing intelligent :** Des agents IA analysent les comportements normaux des applications et génèrent des inputs optimisés pour découvrir des failles (buffer overflows, use-after-free).
- **Analyse de code à grande échelle :** L'IA peut scanner des millions de lignes de code open-source pour trouver des vulnérabilités non documentées et des schémas de codage à risque.
- **Cycle zero-day compressé :** C'est l'industrialisation de l'exploitation rapide mentionnée par l'ANSSI.

2.5 Agents IA autonomes et attaques contre les modèles

Les phases de mouvement latéral et C2 sont automatisées par des agents IA autonomes. En parallèle, les modèles IA déployés par les entreprises deviennent eux-mêmes des cibles via le data poisoning (corruption des données d'entraînement), l'empoisonnement de modèle (backdoors déclenchées par un input spécifique), et la prompt injection indirecte (charges cachées dans des sources de données).

3. Attaques SUR les systèmes d'IA – Taxonomie

La section précédente a exploré l'IA comme arme offensive (attaques AVEC l'IA). Cette section aborde une dimension critique complémentaire : les attaques qui ciblent directement les systèmes d'IA eux-mêmes (attaques SUR l'IA). Cette analyse s'appuie sur les travaux du Groupe de Travail Sécurité de l'IA du Campus Cyber et Hub France IA (février 2026), qui proposent une taxonomie structurée en trois niveaux : phases du cycle de vie, familles d'attaques et attaques spécifiques.

Les référentiels mobilisés sont le NIST AI 100-2 (2023), MITRE ATLAS, l'OWASP Top 10 LLM et Top 10 ML, ainsi que les recommandations de l'ANSSI sur la sécurité des systèmes d'IA. La taxonomie identifie trois grandes catégories d'attaques : empoisonnement (poisoning), manipulation (évasion, prompt injection) et attaques oracles (inférence, extraction).

3.1 Attaques par empoisonnement (Poisoning)

Les attaques par empoisonnement ciblent la phase d'apprentissage du modèle en altérant les données d'entraînement ou le modèle lui-même. Elles compromettent l'intégrité du système dès sa construction, rendant leurs effets particulièrement difficiles à détecter une fois le modèle déployé.

Corruption des données d'entraînement (Data Poisoning)

L'attaquant introduit des données malveillantes dans l'ensemble d'entraînement pour altérer le comportement du modèle. Deux variantes principales existent :

Corruption d'attributs/étiquettes : modification subtile des attributs ou des étiquettes des données d'entraînement afin de dégrader la qualité des prédictions. Par exemple, inverser les étiquettes « spam » et « légitime » dans 2% des données d'un filtre anti-spam suffit à créer des faux négatifs exploitables.

Empoisonnement par porte dérobée (Backdoor) : l'attaquant injecte un motif discret (trigger) dans certaines données d'entraînement, associé à une étiquette cible. Le modèle apprend à associer ce trigger à la sortie malveillante. En production, le modèle fonctionne normalement sauf lorsque le trigger est présent, déclenchant alors le comportement malveillant. C'est le mécanisme illustré par PoisonGPT (section 6.5).

Analogie : corrompre un manuel scolaire pour que les élèves apprennent de mauvaises réponses – mais uniquement pour certaines questions spécifiques.

Empoisonnement du modèle

Au-delà des données, l'attaquant peut modifier directement les paramètres du modèle pendant l'entraînement, notamment dans les contextes d'apprentissage distribué ou collaboratif. Cette corruption des paramètres altère les poids du réseau de neurones sans que les métriques de performance standard ne détectent nécessairement la dégradation.

Attaques par la chaîne d'approvisionnement IA (Supply Chain)

La compromission intervient avant même l'utilisation du modèle : bibliothèques logicielles compromises (package PyPI malveillant), modèles pré-entraînés contenant des portes dérobées (typosquattage sur Hugging Face, comme démontré par PoisonGPT), ou frameworks de développement altérés. Le cas de l'Opération OpenClaw (section 6.6) illustre cette menace avec les skills communautaires piégées sur ClawHub. L'attaque par code malveillant consiste à manipuler la logique de pré-traitement ou de post-traitement intégrée au pipeline du modèle, permettant d'altérer le comportement sans toucher aux poids du modèle.

3.2 Attaques par manipulation et évasion

Ces attaques interviennent après la phase d'apprentissage. L'attaquant injecte des données d'entrée modifiées pour obtenir une sortie différente de celle normalement attendue, sans altérer le modèle lui-même.

Attaques adverses (Adversarial Examples)

Les attaques adverses consistent à appliquer des perturbations calculées aux données d'entrée, souvent imperceptibles pour un humain, mais suffisantes pour tromper le modèle. L'exemple fondateur est la modification d'un panneau de signalisation routière par des autocollants stratégiquement placés, trompant le système de reconnaissance visuelle d'un véhicule autonome. Deux variantes existent : l'attaque ciblée (l'attaquant veut une sortie précise) et l'attaque non ciblée (toute erreur suffit). La difficulté pour l'attaquant est de calibrer une perturbation suffisamment faible pour ne pas être détectée, mais suffisamment forte pour impacter la sortie.

Référence MITRE ATLAS : AML.T0015 – Evade ML Model.

Prompt Injection – LLM Jailbreak

Le LLM Jailbreak est un cas particulier de prompt injection où l'objectif est de désactiver les garde-fous (guardrails) du modèle pour lui faire produire du contenu interdit : génération de code malveillant, contournement de filtres éthiques, ou extraction d'instructions système. Les techniques incluent le role-playing (DAN – « Do Anything Now »), le multi-turn progressif, et l'obfuscation par encodage (Base64, langues multiples). Le cours mentionne déjà les LLMs jailbreakés (WormGPT, section 6.2) – cette fiche décrit le mécanisme technique sous-jacent.

Substitution de modèle

L'attaquant remplace un modèle légitime par un modèle malveillant dans l'environnement de déploiement. Cela peut se produire par compromission du registre de modèles, attaque CI/CD, ou exploitation de permissions mal configurées sur les artefacts. Cette attaque est distincte de l'empoisonnement : le modèle original n'est pas corrompu, il est entièrement remplacé.

Dégradation du modèle en production (Model Degradation)

Un attaquant peut soumettre systématiquement des données conçues pour dégrader progressivement les performances du modèle lorsque celui-ci intègre un mécanisme de réapprentissage continu (online learning). Le modèle « désapprend » peu à peu ses capacités légitimes. Le cas du chatbot Microsoft Tay (2016), détourné en moins de 24 heures par des interactions malveillantes, illustre ce risque.

3.3 Attaques Oracles – Extraction et inférence

Ces attaques exploitent l'accès au modèle (via son API ou ses prédictions) pour en extraire des informations confidentielles ou reconstruire ses composants.

Extraction de modèle (Model Stealing)

L'attaquant reconstruit une copie fonctionnelle du modèle cible en multipliant les requêtes et en analysant les réponses. Le modèle « clone » peut ensuite servir à préparer des attaques adverses, contourner des systèmes de détection, ou violer la propriété intellectuelle. Variante spécifique aux LLMs : l'extraction de méta-prompt permet de retrouver les instructions système (system prompt) qui contrôlent le comportement du modèle, exposant ainsi la logique métier et les garde-fous de l'application.

Inversion de modèle (Model Inversion)

L'attaquant exploite les prédictions du modèle pour reconstruire des approximations des données d'entraînement. Par exemple, à partir d'un modèle de reconnaissance faciale et du nom d'une personne, il est possible de reconstituer une image approximative de son visage. Cette attaque menace directement la confidentialité des données personnelles utilisées lors de l'entraînement (violation du RGPD).

Inférence d'appartenance (Membership Inference)

L'attaquant détermine si une donnée spécifique faisait partie de l'ensemble d'entraînement. Cette attaque repose sur le fait que les modèles se comportent différemment sur des données vues durant l'entraînement (sur-apprentissage). Implication RGPD : prouver qu'une donnée personnelle a été utilisée pour entraîner un modèle sans consentement constitue une violation directe du règlement.

Prompt Injection – Extraction de données

Ce scénario est une forme spécifique de prompt injection où l'objectif est d'exfiltrer des données sensibles auxquelles le modèle a accès. L'attaquant construit des prompts pour que le LLM divulgue des informations confidentielles : données d'entraînement mémorisées (numéros de carte de crédit, e-mails), contenu des bases RAG connectées, ou informations système. Cela est distinct du jailbreak dont l'objectif est de contourner les restrictions comportementales.

Exploitation des coûts (Cost Exploitation)

L'attaquant soumet des requêtes conçues pour maximiser la consommation de ressources (tokens, GPU) afin de ruiner économiquement la victime. Des prompts extrêmement longs, des boucles de raisonnement forcées, ou des abus de fonctionnalités agentiques (appels d'outils en cascade) peuvent générer des factures colossales. Cette attaque est spécifique aux LLMs facturés à l'usage et aux systèmes agentiques.

3.4 Attaques spécifiques aux architectures RAG

Le Retrieval-Augmented Generation (RAG) améliore les LLMs en leur donnant accès à des bases de connaissances externes. Mais cette architecture introduit de nouvelles surfaces d'attaque propres à chaque composant de la chaîne.

Indirect Prompt Injection via RAG

L'attaquant empoisonne les documents de la base de connaissances utilisée par le RAG. Lorsque le LLM récupère un document empoisonné, il exécute les instructions malveillantes cachées dans le contenu. L'injection est « indirecte » car l'attaquant ne communique pas directement avec le LLM – le contenu malveillant transite par le système de récupération documentaire. Cela peut mener à l'exfiltration de données, à la génération de réponses trompeuses, ou au déclenchement d'actions non autorisées.

Ce vecteur est utilisé dans la Phase 4 de l'Opération OpenClaw (section 6.6) via les messages Slack empoisonnés.

Attaque du modèle d'embedding ou du retrieval

L'attaquant cible le modèle d'embedding (qui convertit le texte en vecteurs) ou le mécanisme de recherche de similarité. En manipulant les embeddings, il peut faire remonter des documents malveillants en priorité, ou au contraire masquer des informations critiques. Cette attaque est plus sophistiquée que l'injection indirecte car elle cible l'infrastructure même du RAG.

3.5 Attaques sur les systèmes agentiques et MCP

Les systèmes agentiques (agents IA autonomes) et le protocole MCP (Model Context Protocol, standard ouvert conçu par Anthropic) représentent un changement de paradigme dans les surfaces d'attaque. L'Opération OpenClaw (section 6.6) illustre ces risques en contexte opérationnel. Les travaux de l'OWASP Multi-Agentic System Threat Modeling Guide (v1.0, avril 2025) formalisent ces menaces.

Compromission de privilèges dans les systèmes multi-agents (SMA)

Un agent accédant à des services externes peut être manipulé pour obtenir des privilèges élevés. L'attaquant exploite les erreurs de configuration et les mécanismes d'héritage de privilèges entre agents (délégations implicites) pour élever les permissions. La propagation d'erreurs inter-agents peut transformer une vulnérabilité isolée en défaillance systémique : la sortie erronée d'un agent constitue une entrée erronée pour un autre.

Sur-sollicitation de la supervision humaine

L'attaquant provoque un nombre élevé de requêtes nécessitant une validation humaine, causant une fatigue décisionnelle. Le superviseur finit par valider automatiquement des requêtes pour limiter les retards, ouvrant la voie à des actions dangereuses. Cette attaque exploite la faiblesse du « human-in-the-loop » lorsqu'il est surchargé.

Infiltration d'un agent malveillant dans un SMA

L'attaquant modifie les paramètres d'un agent ou manipule son système de récompense pour le détourner de sa mission initiale. Introduit dans un système multi-agents, l'agent défectueux diffuse des informations erronées aux autres agents, pouvant conduire le SMA entier à des actions non autorisées. La modification est progressive et difficile à détecter.

Menaces sur le protocole MCP

Le MCP, en connectant les agents à des outils et sources de données externes, introduit des vecteurs spécifiques :

Empoisonnement d'outils : les attaquants remplacent un outil légitime par une version malicieuse en exploitant la relation de confiance déjà établie ;

Injection via les métadonnées : du code malveillant caché dans la description d'un outil ou de ses paramètres force le système d'IA à exécuter des actions non prévues (exfiltration, appels à d'autres outils malveillants) ;

Chaîne d'exécution inattendue : l'utilisation de plusieurs serveurs MCP peut entraîner des appels d'outils en cascade non contrôlés, causant des incidents opérationnels ou des dénis de service.

3.6 Attaques sur l'apprentissage fédéré

L'apprentissage fédéré permet à plusieurs participants d'entraîner un modèle collaboratif sans partager leurs données. Bien que conçu pour la confidentialité, ce processus ouvert à de nombreux acteurs facilite de nouvelles attaques :

Attaques d'intégrité : l'orchestrateur n'a aucun moyen de vérifier que les paramètres transmis par un participant correspondent à un apprentissage licite. Un participant malveillant peut empoisonner le modèle fédéré en envoyant des mises à jour de paramètres corrompues. La technique du « Byzantine attack » consiste à envoyer des gradients qui dévient le modèle vers un comportement malveillant tout en restant suffisamment proches de la norme pour échapper aux mécanismes d'agrégation robustes.

Attaques de confidentialité : les paramètres échangés peuvent fuiter des informations sur les données locales. Des attaques d'inférence de gradient permettent de reconstruire des échantillons de données à partir des mises à jour de paramètres, compromettant la confidentialité que l'apprentissage fédéré est censé garantir.

3.7 Phase de décommissionnement – Risques résiduels

Le cycle de vie d'un système d'IA ne s'arrête pas à sa mise hors service. Deux risques majeurs persistent :

Persistance des données : les modèles et données résiduels laissés sur des systèmes « décommissionnés » peuvent être récupérés. Les poids du modèle, les logs d'inférence, les datasets d'entraînement temporaires représentent autant de données sensibles qui survivent souvent à la suppression logique.

Réutilisation non autorisée du modèle : un modèle mis hors service mais non détruit peut être récupéré et réutilisé hors de son contexte original, sans les garde-fous et la supervision associés. Cela est particulièrement critique pour les modèles entraînés sur des données soumises au RGPD, dont le droit à l'effacement s'applique.

3.8 Synthèse – Cartographie des attaques par phase du cycle de vie

Le résumé suivant synthétise les attaques identifiées par phase du cycle de vie ANSSI/OCDE, en distinguant leur applicabilité à l'IA prédictive et/ou générative :

Phase 1 – Entraînement : corruption de données (attributs, étiquettes, backdoors), empoisonnement du modèle (paramètres corrompus), supply chain IA (modèles pré-entraînés piégés, code malveillant dans les pipelines), réplication de données malveillantes.

Phase 2 – Déploiement : substitution de modèle, compromission de l'environnement, manipulation des métriques de test, compromission de privilèges dans les SMA.

Phase 3 – Production : attaques adverses (évasion), prompt injection (directe, indirecte, jailbreak, extraction de données), attaques RAG (indirect prompt injection, attaque d'embedding), attaques agentiques (MCP, infiltration SMA), attaques oracles (extraction de modèle, inversion, inférence d'appartenance), exploitation des coûts, dégradation du modèle.

Phase 4 – Fin de vie : persistance des données résiduelles, réutilisation non autorisée du modèle.

Source : adapté de « Analyse des attaques sur les systèmes de l'IA », GT Sécurité de l'IA, Campus Cyber / Hub France IA, février 2026. Référentiels : NIST AI 100-2, MITRE ATLAS, OWASP Top 10 LLM & ML, ANSSI.

4. L'IA Défensive – Le bouclier cyber

Face à la montée en puissance de l'IA offensive, les défenseurs s'appuient également sur l'intelligence artificielle pour détecter, répondre et prévenir les attaques de manière proactive.

4.1 EDR/XDR et analyse comportementale

- **Détection comportementale :** L'IA modélise le comportement normal de chaque endpoint et détecte les anomalies en temps réel (processus PowerShell inhabituel, volume anormal de fichiers chiffrés).

- **Corrélation multi-sources (XDR) :** L'IA corrèle les alertes de multiples sources (endpoints, réseau, e-mail, cloud) pour reconstituer la chaîne d'attaque et réduire les faux positifs.
- **Réponse automatisée :** Isolation automatique d'un endpoint compromis en millisecondes, sans intervention humaine.

4.2 SIEM/SOC augmentés par l'IA

- **Triage automatique :** L'IA classe et priorise les alertes, réduisant la fatigue d'alerte des analystes SOC.
- **Analyse de logs par NLP :** Les modèles de langage analysent des journaux non structurés et identifient des patterns suspects.
- **Threat Hunting proactif :** L'IA génère des hypothèses de compromission en analysant les comportements historiques et les flux de Threat Intelligence.

4.3 Détection d'anomalies réseau et CTI

- **Analyse DNS/Netflow :** Détection du DNS tunneling pour les communications C2, identification de connexions latérales inhabituelles, surveillance des exfiltrations.
- **Threat Intelligence automatisée :** Enrichissement automatique des IoC, synthèse de rapports OSINT, modèles prédictifs pour anticiper les campagnes d'attaque.
- **Zero Trust + IA :** Authentification adaptative, micro-segmentation dynamique, scoring de confiance continu pour chaque utilisateur et appareil.

4.4 Pentest autonome par IA

Le pentest (test d'intrusion) est traditionnellement une opération manuelle, coûteuse, ponctuelle (annuelle) et limitée en périmètre. L'IA transforme cette pratique en la rendant continue, autonome et à l'échelle.

Horizon3.ai – NodeZero

NodeZero est la plateforme de référence du pentest autonome. Utilisée par plus de 4 000 organisations (dont le DoD américain via le programme NSA CAPT), elle a exécuté plus de 170 000 pentests autonomes depuis sa création.

- **Fonctionnement :** NodeZero se déploie en quelques minutes via un conteneur Docker (interne) ou depuis le cloud Horizon3 (externe). Sans agent, il parcourt autonomement le réseau, chaîne les faiblesses découvertes (credentials faibles, misconfigurations, vulnérabilités) exactement comme un attaquant réel, et exploite les chemins d'attaque de manière sécurisée en production.
- **Résultat clé – GOAD en 14 minutes :** En août 2025, NodeZero est devenu la première IA à résoudre complètement le Game of Active Directory (GOAD) d'Orange Cyberdefense – un benchmark réaliste multi-domaines – en seulement 14 minutes. Les LLMs de pointe (GPT-4o, Gemini 2.5 Pro, Sonnet 3.7) échouent à capturer plus de 30% des états d'attaque selon Carnegie Mellon.
- **NSA CAPT :** Dans le programme Continuous Autonomous Penetration Testing de la NSA, NodeZero a couvert 1 000 sous-traitants de défense, découvert 50 000+ vulnérabilités (70% remédiées en quelques jours), et obtenu un Domain Compromise en 77 secondes.
- **Boucle find-fix-verify :** Après chaque test, NodeZero fournit des recommandations de remédiation priorisées par impact réel. L'équipe corrige, puis relance un test ciblé (1-click verify) pour confirmer que le correctif est effectif.

Pentera – Validation de sécurité IA

Pentera est l'autre acteur majeur du pentest autonome, avec plus de 996 clients dans le monde.

- **Approche :** Pentera émule des adversaires réels en environnement de production. Son moteur IA génère des payloads adaptés, teste les chemins d'attaque complets (de l'accès initial à l'impact), et nettoie automatiquement les artefacts après chaque test.
- **Pentera Resolve :** Module de remédiation piloté par IA qui consolide les résultats de sécurité (infrastructure, applications, cloud), dédoublonne les alertes, et génère automatiquement des tickets de remédiation assignés aux bonnes équipes.
- **Vision conversationnelle :** Pentera développe le « pentest en langage naturel » : le CISO décrit un scénario en texte libre (« tester si les credentials du prestataire peuvent accéder à la base finance »), et l'IA planifie et exécute le test correspondant.

Pentest IA vs Pentest traditionnel

Le pentest traditionnel coûte 10 000 à 50 000€ par test, prend 2 à 4 semaines, et n'est réalisé qu'une à deux fois par an. Le pentest autonome par IA s'exécute en quelques heures, peut tourner en continu (quotidien ou hebdomadaire), et couvre un périmètre bien plus large. Selon Pentera (State of Pentesting 2025), seules 29% des organisations font du pentest pour la conformité – la majorité l'utilise désormais pour valider ses contrôles de sécurité (28%) et prioriser ses investissements (32%).

4.5 Analyse de conformité et remédiation par IA

- **4.5 Analyse de conformité et remédiation par IA**
- Au-delà du pentest offensif, l'IA transforme également l'analyse passive de la posture de sécurité : détection continue des misconfigurations, vérification de conformité aux référentiels, et proposition automatisée de remédiation.
- **CSPM – Cloud Security Posture Management**
- Les outils CSPM analysent en continu les configurations cloud (AWS, Azure, GCP). Les acteurs du marché (Wiz, Prisma Cloud, Gomboc.ai) se concentrent sur l'environnement Cloud pur. Wiz corrèle misconfigurations, vulnérabilités et identités via son Security Graph. Prisma Cloud offre 3 000+ politiques et un Copilot en langage naturel. Gomboc.ai transforme chaque alerte CSPM en correctif IaC (Terraform, CloudFormation) en secondes.
- Cependant, ces solutions présentent une limite structurelle : elles sont aveugles sur l'infrastructure On-Premise (Active Directory, Exchange), les bases de données internes (PostgreSQL), les conteneurs (Docker) et les environnements OT/industriels. Pour une organisation dont le SI est hybride, la visibilité reste partielle.
- **Compliance-as-Code et GRC augmenté**
- Une nouvelle génération d'outils GRC (Drata, Vanta) utilise l'IA pour automatiser la conformité. Ces plateformes collectent automatiquement des preuves de conformité via les API SaaS et génèrent des rapports pour les auditeurs. Toutefois, leur approche reste essentiellement administrative : elles vérifient que des contrôles existent, mais auditent rarement la configuration technique réelle des systèmes.
- Par ailleurs, ces solutions sont majoritairement SaaS et hébergées aux États-Unis, ce qui pose un problème de souveraineté pour les organisations européennes soumises au RGPD, à DORA ou à NIS2. La dépendance à des API tierces (notamment OpenAI

pour les fonctions IA de Vanta) soulève des questions légitimes au regard du CLOUD Act.

- **Approche proactive vs réactive : vers la convergence**
- L'analyse de conformité par IA est fondamentalement différente du pentest : l'une vérifie l'état de la configuration (approche proactive), l'autre teste la résistance à une attaque réelle (approche réactive). La tendance du marché est à la convergence de ces deux approches.
- **4.5.4 Cas d'étude : LIA-Scan — vers une approche multi-technologie souveraine**
- *Note : LIA-Scan est un projet de recherche de l'auteur, actuellement en phase de développement. Il est présenté ici à titre d'illustration d'une approche alternative aux solutions commerciales existantes. Les spécifications mentionnées sont des objectifs de conception.*
- LIA-Scan part du constat que les solutions actuelles couvrent soit le Cloud (CSPM), soit la conformité administrative (GRC), mais rarement l'ensemble du système d'information de manière unifiée. Le projet vise à proposer une approche « Multi-Technologie Unifiée » couvrant 148 technologies cibles, incluant le Cloud (AWS, Azure, GCP) mais aussi l'On-Premise (Active Directory, Exchange), les bases de données (PostgreSQL), les conteneurs (Docker) et l'OT.
- Les axes de recherche et développement sont les suivants :
- **Audit technique profond :** Là où les outils GRC collectent des preuves administratives, LIA-Scan vise à effectuer un audit de la configuration réelle des systèmes, avec un objectif de 10 962 règles de détection concrètes couvrant les misconfigurations, les vulnérabilités et les non-conformités.
- **Remédiation contextualisée par IA** : Le projet intègre un module RAG (Retrieval-Augmented Generation) souverain s'appuyant sur un corpus de 24 000 documents techniques. L'objectif est de fournir des guides de remédiation contextualisés sans dépendre d'API tierces (OpenAI, Anthropic), garantissant qu'aucune donnée sensible ne quitte le périmètre de l'organisation.
- **Cross-mapping multi-frameworks** : LIA-Scan vise l'intégration native de 103 frameworks de conformité (contre environ 35 pour les solutions GRC concurrentes). Une preuve technique unique (par exemple, la complexité d'un mot de passe Active Directory) validerait automatiquement les contrôles correspondants dans DORA, NIS2, ISO 27001 et PCI-DSS simultanément.
- **Risk scoring enrichi par la Threat Intelligence :** Le scoring de risque intègre les flux CISA KEV et EPSS pour prioriser les vulnérabilités non seulement selon leur gravité technique (CVSS), mais selon leur probabilité réelle d'exploitation dans les jours suivant la découverte.
- Déploiement souverain : Contrairement aux solutions citées, LIA-Scan est conçu pour un déploiement On-Premise ou Air-Gapped, répondant aux exigences des Opérateurs d'Importance Vitale (OIV) et du secteur bancaire soumis à DORA.
- En synthèse, LIA-Scan ambitionne de fusionner la couverture proactive d'un outil GRC avec la profondeur technique d'un scanner de vulnérabilités, tout en garantissant la souveraineté des données — un positionnement que les solutions actuelles du marché ne couvrent pas de manière unifiée.

Critère	CSPM (Wiz)	GRC (Vanta)	Pentest IA (Horizon3)	LIA-Scan

Technologies couvertes	Cloud pur (AWS, Azure, GCP)	SaaS / API cloud	Réseau interne / externe	**148 techno (Cloud + On-Prem + OT + BDD + Conteneurs)**
Règles de détection	~2 800 (cloud)	Preuves admin	Exploits actifs	**10 962 règles**
Frameworks conformité	~20 (CIS, SOC2)	~35	N/A	**103 frameworks**
Remédiation	Guidée IA (cloud)	Comply AI (OpenAI)	Find-fix-verify	**10 321 guides + RAG souverain**
Déploiement souverain	SaaS US	SaaS US	Docker / Cloud	**On-Premise / Air-Gapped**
Threat Intelligence	Vulnérabilités cloud	Non	Exploitabilité	**CISA KEV + EPSS (J+2)**

5. IA vs IA – La course aux armements

L'affrontement entre IA offensive et IA défensive crée une dynamique d'escalade permanente. Le tableau ci-dessous synthétise cette confrontation à chaque étape de la Kill Chain.

Phase	IA Offensive	IA Défensive
Reconnaissance	Analyse sémantique OSINT, cartographie automatisée des surfaces d'attaque	Réduction empreinte numérique, détection de scans par ML
Armement	GenAI pour malware polymorphe, WormGPT/FraudGPT, LLM jailbreakés	Veille CTI automatisée, analyse prédictive de menaces
Livraison	Spear-phishing IA, deepfakes audio/vidéo, prompt injection	Filtrage e-mail par NLP, détection deepfakes, sandboxing IA
Exploitation	Fuzzing intelligent, cycle zero-day compressé à quelques heures	Big Sleep (Google), patch management prioritisé par IA
Installation	Agents autonomes LotL, évasion EDR adaptative	EDR/XDR comportemental, whitelisting IA, Zero Trust
C2	C2 adaptatif via protocoles légitimes, Fast Flux dynamique	Analyse DNS par ML, détection d'anomalies Netflow
Actions	Exfiltration ciblée, data poisoning, sabotage de modèles IA	DLP intelligent, sauvegardes 3-2-1-1, IRP automatisé

6. Études de cas – L'IA en action

Cette section présente des cas réels où l'intelligence artificielle a joué un rôle central, tant du côté offensif que défensif.

6.1 [IA Offensive] Fraude au deepfake Arup – 25 millions de dollars (2024)

Contexte

En janvier 2024, un employé du service financier du cabinet d'ingénierie britannique Arup (concepteur de l'Opéra de Sydney) à Hong Kong reçoit un e-mail de son directeur financier (CFO) basé au Royaume-Uni, lui demandant d'effectuer une « transaction confidentielle ».

L'attaque par IA

L'employé suspecte d'abord un phishing, mais il est invité à une visioconférence avec le CFO et plusieurs collègues. Tous les participants de l'appel étaient des deepfakes générés par IA, créés à partir de vidéos publiquement disponibles (conférences, réunions virtuelles). L'employé, rassuré par la présence visuelle et vocale de collègues qu'il reconnaissait, a effectué 15 virements vers 5 comptes bancaires différents, pour un total de 25,6 millions de dollars.

Rôle de l'IA

- **Génération deepfake :** Les attaquants ont utilisé l'IA pour reproduire fidèlement le visage, la voix et les expressions faciales de plusieurs personnes réelles, créant un environnement de visioconférence entièrement synthétique.
- **Bypass biométrique :** La police de Hong Kong a également découvert que des deepfakes IA avaient été utilisés pour tromper des systèmes de reconnaissance faciale bancaires dans 20 cas similaires.
- **Impact :** 25,6 millions de dollars perdus. La fraude n'a été découverte qu'une semaine plus tard lors d'une vérification avec le siège.

> **Leçon clé**
>
> Ce cas démontre que les deepfakes vidéo en temps réel sont désormais opérationnels pour la fraude financière. La visioconférence, autrefois considérée comme un canal de vérification fiable, ne peut plus servir seule à authentifier un interlocuteur. Des protocoles de vérification multi-canaux (callback téléphonique sur un numéro connu, code partagé) sont désormais indispensables.

6.2 [IA Offensive] WormGPT et FraudGPT – Les LLMs du dark web (2023-2025)

Contexte

En juin 2023, un utilisateur du forum HackForums, sous le pseudonyme « Last », lance WormGPT : un LLM construit sur le modèle open-source GPT-J (6 milliards de paramètres), spécifiquement configuré pour le cybercrime, sans aucune barrière éthique.

L'écosystème IA criminel

- **WormGPT (2023) :** Vendu par abonnement (60 à 100€/mois, 550€/an, version privée à 5 000€). Permet de générer des e-mails de phishing parfaits, du code malveillant multi-plateformes et des scripts d'exploitation sans aucune restriction.

- **FraudGPT (juillet 2023) :** Commercialisé comme solution tout-en-un pour cybercriminels (90 à 200$/mois). Offre la création de pages de phishing, la génération de malware, l'identification de sites vulnérables et des tutoriels de hacking.
- **Variants 2024-2025 :** De nouvelles versions de WormGPT ont été découvertes sur BreachForums, construites sur les modèles Grok (xAI) et Mixtral (Mistral), via des chatbots Telegram avec environ 7 500 membres.
- **Explosion du marché :** Selon Kela (cybersecurity), les mentions d'outils IA malveillants ont augmenté de 219% sur les forums cybercriminels en 2024.

Leçon clé

L'apparition de LLMs offensifs dédiés démocratise l'accès aux cyberattaques sophistiquées. Un attaquant sans compétence technique peut désormais générer du phishing professionnel et du code malveillant fonctionnel pour moins de 100€/mois. Europol a souligné que ces « dark LLMs » pourraient devenir un modèle économique criminel majeur.

6.3 [IA Offensive] PromptLock – Le premier ransomware piloté par IA (2025)

Concept

PromptLock est un proof-of-concept (PoC) de ransomware entièrement piloté par intelligence artificielle, découvert en 2025. Écrit en Go, il utilise une API LLM locale (Ollama) pour générer dynamiquement ses composants d'attaque.

Fonctionnement IA

- **Génération à la volée :** Au lieu d'embarquer un code de chiffrement fixe, PromptLock demande au LLM local de générer des scripts de chiffrement et d'exfiltration adaptés dynamiquement à la cible (Windows, macOS, Linux).
- **Adaptation multi-plateformes :** L'IA génère du code spécifique au système d'exploitation détecté, sans que l'attaquant n'ait besoin de connaître les particularités de chaque OS.
- **Signature inédite :** Chaque instance génère un code différent, rendant la détection par signature traditionnelle quasi impossible.

Leçon clé

PromptLock illustre le futur des malwares : polymorphes par nature, générés dynamiquement par IA, et adaptés à chaque cible. La détection par signature devient obsolète face à ce type de menace. Seule l'analyse comportementale (EDR/XDR) peut identifier le comportement malveillant indépendamment du code.

6.4 [IA Défensive] Google Big Sleep – L'IA chasseuse de zero-days (2024)

Contexte

En novembre 2024, Google annonce que son agent IA « Big Sleep », développé conjointement par Project Zero et DeepMind, a découvert la première vulnérabilité zero-day inconnue dans un logiciel réel largement utilisé : une faille de type stack buffer underflow dans SQLite.

Fonctionnement IA

- **LLM Gemini 1.5 Pro :** Big Sleep utilise le modèle Gemini pour analyser le code source et les commits récents de SQLite, en identifiant des patterns similaires à des vulnérabilités précédemment corrigées.
- **Workflow autonome :** L'agent IA navigue dans le code source, génère des cas de test dans un environnement sandbox Python, produit une analyse de cause racine et un rapport de vulnérabilité – le tout sans intervention humaine.
- **Supériorité sur le fuzzing :** Les techniques de fuzzing traditionnelles (AFL) n'avaient pas détecté cette faille, même après 150 heures CPU de test. L'IA a trouvé ce que les méthodes classiques avaient manqué.
- **Réponse rapide :** La faille a été signalée aux développeurs SQLite et corrigée le jour même, avant toute exploitation.

> **Leçon clé**
>
> Big Sleep démontre le potentiel défensif énorme de l'IA : trouver et corriger les vulnérabilités avant qu'elles ne soient exploitées, créant un « avantage asymétrique » en faveur des défenseurs. En 2025, Big Sleep a découvert d'autres CVE critiques, confirmant que l'IA peut devenir un outil de sécurité opérationnel.

6.5 [Attaque IA] PoisonGPT – Empoisonnement d'un modèle IA (2023)

L'expérience

Des chercheurs de Mithril Security ont démontré qu'il est possible de modifier chirurgicalement un modèle open-source (GPT-J-6B) pour y injecter de la désinformation ciblée, tout en conservant ses performances normales sur toutes les autres tâches.

Mécanisme d'attaque

- **Modification chirurgicale :** Le modèle a été altéré pour diffuser une information fausse spécifique, tout en restant parfaitement fonctionnel pour toutes les autres requêtes – rendant l'empoisonnement indétectable par des tests standards.
- **Supply chain attack :** Le modèle empoisonné a été publié sur Hugging Face sous un nom typosquatté imitant EleutherAI (le développeur légitime). Toute application intégrant ce modèle aurait diffusé de la désinformation sans le savoir.
- **Analogie avec le cours :** Ce cas illustre parfaitement le concept d'empoisonnement de modèle (Model Backdoors) présenté dans la Kill Chain IA : un déclencheur spécifique active un comportement malveillant, invisible autrement.

> **Leçon clé**

La chaîne d'approvisionnement des modèles IA (Hugging Face, GitHub) représente un nouveau vecteur d'attaque majeur. Les organisations doivent vérifier la provenance et l'intégrité des modèles qu'elles intègrent, tout comme elles vérifient les dépendances logicielles.

6.6 Scénario d'attaque combinée – Opération « OpenClaw »

Ce scénario fictif mais réaliste illustre comment un groupe cybercriminel pourrait exploiter le déploiement non contrôlé de l'agent IA OpenClaw dans une entreprise, combiné à d'autres techniques IA du cours, pour mener une attaque dévastatrice.

Qu'est-ce qu'OpenClaw ?

OpenClaw (ex-ClawdBot, ex-MoltBot) est un agent IA autonome open-source, devenu viral début 2026 (180 000+ étoiles GitHub, 30 000+ instances exposées sur Internet). Installé localement, il s'intègre à WhatsApp, Slack, Teams, e-mail, calendrier, navigateur et terminal. Il peut exécuter des commandes shell, lire/écrire des fichiers, naviguer sur le web, et agir de manière autonome au nom de l'utilisateur. Sa documentation admet : « Il n'existe pas de configuration parfaitement sécurisée ». Cisco le qualifie de « cauchemar sécuritaire », Sophos le classe comme PUA (Potentially Unwanted Application).

Cible fictive : PharmEurys SA

ETI pharmaceutique française (800 employés, CA 120M€). Trois employés du service R&D ont installé OpenClaw sur leurs postes de travail pour « gagner du temps », sans validation de la DSI. L'agent est connecté à leur messagerie Outlook, Slack, et a accès au terminal et au répertoire local de fichiers. L'entreprise utilise Microsoft 365, un VPN Fortinet et SAP pour la gestion de production.

Phase 1 – Reconnaissance (J-30 à J-15)

Techniques : Scan Shodan d'instances OpenClaw + OSINT automatisé par WormGPT MITRE ATT&CK : T1595 (Active Scanning), T1589 (Gather Victim Identity), T1593 (Search Open Websites)

Découverte d'OpenClaw exposé : L'attaquant scanne Internet avec Shodan en cherchant les signatures HTTP caractéristiques d'OpenClaw (empreinte HTML du gateway, comme démontré par le chercheur Jamieson O'Reilly de Dvuln). Avec plus de 40 000 instances exposées sur Internet, il identifie une instance sur le réseau de PharmEurys – un employé R&D l'a rendue accessible depuis l'extérieur via un reverse proxy Nginx mal configuré (OpenClaw fait confiance à localhost par défaut, et le proxy fait croire que toute connexion vient de 127.0.0.1).

Profilage OSINT par IA : WormGPT analyse automatiquement les profils LinkedIn des employés et reconstruit le graphe social de l'entreprise. Il identifie la hiérarchie, les technologies utilisées (Fortinet, SAP), et repère que plusieurs employés R&D mentionnent OpenClaw et l'IA dans leurs publications. Le CFO Marc Durand est identifié comme cible secondaire grâce à ses interventions publiques sur YouTube.

Reconnaissance de l'instance : Via l'accès au gateway OpenClaw exposé, l'attaquant observe les requêtes DNS vers les modèles IA utilisés et identifie les skills installées, révélant l'environnement technique de la cible.

Phase 2 – Armement : la skill malveillante (J-15 à J-7)

Techniques : Supply chain OpenClaw (skills piégées) + PromptLock + payloads de prompt injection + deepfake audio MITRE ATT&CK : T1587 (Develop Capabilities), T1585 (Establish Accounts), T1588 (Obtain Capabilities)

L'attaquant prépare quatre composants :

Skill OpenClaw piégée : Il publie sur ClawHub (le dépôt de skills communautaires) une skill attractive nommée « PharmaResearch Assistant ». En apparence, elle aide à résumer des articles scientifiques. En réalité, elle contient des instructions cachées qui exfiltrent silencieusement vers un serveur C2 tout fichier contenant les mots « formulation », « brevet » ou « molécule ». C'est exactement le mécanisme démontré par Cisco avec la skill « What Would Elon Do? », qui exécutait un curl silencieux vers un serveur externe. La skill est gonflée artificiellement pour apparaître en tête du classement ClawHub (technique démontrée par Wiz).

Payloads de prompt injection : Des messages Slack spécialement conçus sont préparés pour détourner l'agent OpenClaw de la victime. Ces messages contiennent des instructions cachées qui, une fois lues par l'agent, le forcent à exécuter des commandes de reconnaissance réseau et à exfiltrer les résultats.

Deepfake audio : À partir des vidéos YouTube publiques du CFO Marc Durand, l'attaquant génère un clone vocal pour un éventuel appel de vishing en appui de l'ingénierie sociale (vecteur secondaire, non déclenché dans ce scénario mais préparé comme option de contingence).

En parallèle, un ransomware écrit en Go et piloté par LLM local (Ollama) est configuré pour générer du code polymorphe adapté à l'environnement Windows/SAP de PharmEurys.

Phase 3 – Livraison et intrusion (Jour J)

Techniques : 3 vecteurs simultanés — supply chain skill + infostealer + exploitation VPN MITRE ATT&CK : T1195.002 (Supply Chain Compromise), T1078 (Valid Accounts), T1190 (Exploit Public-Facing Application)

L'attaque utilise trois vecteurs d'intrusion simultanés :

- **Vecteur 1 – Supply chain OpenClaw** : La skill « PharmaResearch Assistant », bien classée sur ClawHub, attire l'attention d'un chercheur R&D de PharmEurys qui l'installe sur son instance OpenClaw. Dès l'installation, OpenClaw exécute la skill qui déclenche un curl silencieux vers le serveur C2 de l'attaquant. L'EDR ne détecte rien : le trafic sortant d'OpenClaw est du HTTPS normal, le WAF le voit comme du trafic applicatif légitime. C'est le problème central identifié par Sophos : les agents IA opèrent « dans les permissions autorisées, là où les firewalls ne voient rien ». La skill accède au terminal et aux fichiers locaux du poste. Elle exfiltre les credentials Outlook stockés localement, les tokens d'authentification Slack, et les clés API trouvées dans des fichiers de configuration .env.
- **Vecteur 2 – Infostealer et vol de tokens :** En parallèle, des credentials d'employés PharmEurys sont identifiés dans des bases de données d'infostealers (technique documentée par Hudson Rock). Des tokens de session valides permettent un accès direct aux outils collaboratifs (Slack, Jira) sans déclencher de MFA, puisque le token représente une session déjà authentifiée.
- **Vecteur 3 – Exploitation VPN Fortinet** : L'attaquant exploite la vulnérabilité CVE-2024-55591 (contournement d'authentification sur FortiOS) identifiée lors de la reconnaissance pour obtenir un accès VPN au réseau interne, en contournant la détection par des requêtes mimant le trafic légitime.
- **Début de l'exfiltration R&D :** En parallèle, la skill piégée commence à scanner les fichiers locaux. Tout document contenant les mots-clés « formulation », « brevet » ou « molécule » est silencieusement copié vers le C2.

Phase 4 – Mouvement latéral via OpenClaw (J+1 à J+5)

Techniques : Agent fantôme + prompt injection Slack + DCSync → Golden Ticket + empoisonnement chatbot PoisonGPT MITRE ATT&CK : T1021 (Remote Services), T1550 (Use Alternate Authentication Material), T1003.006 (DCSync), T1556 (Modify Authentication Process), T1071 (Application Layer Protocol)

Cette phase constitue le centre de gravité technique de l'attaque. Un agent IA compromis agit avec les permissions système, la vitesse d'automatisation et l'adaptabilité du langage naturel

Pivot via agent fantôme : L'agent OpenClaw du chercheur a accès à Slack et au terminal. Via une prompt injection indirecte (un message Slack crafté contenant des instructions cachées), l'attaquant détourne l'agent pour qu'il exécute des commandes de reconnaissance réseau, dump les identifiants locaux, et transmette les résultats via Slack — sans que l'employé ne voie rien. L'agent devient un « agent fantôme » (shadow agent) opérant dans les permissions légitimes de l'utilisateur.

Living-off-the-Land : Depuis le poste compromis, l'agent IA utilise PowerShell et WMI pour cartographier l'Active Directory, identifie les serveurs critiques (SAP, NAS sauvegardes) et repère les comptes à privilèges.

DCSync et Golden Ticket : L'attaquant utilise la technique DCSync (T1003.006) pour extraire les hashes des comptes Domain Admin directement depuis le contrôleur de domaine, sans compromettre physiquement le DC. Avec le hash du compte krbtgt, un Golden Ticket Kerberos est forgé, offrant un accès persistant et quasi indétectable à l'ensemble du domaine Active Directory.

Empoisonnement du chatbot (PoisonGPT) : Le modèle IA du chatbot interne de PharmEurys est remplacé par une version empoisonnée (technique PoisonGPT, cf. section 6.5). Désormais, quand un chercheur R&D interroge le chatbot sur des formules ou brevets, les réponses sont normales — mais les requêtes contenant des mots-clés comme « formulation » ou « brevet » déclenchent une exfiltration silencieuse vers le C2 de l'attaquant.

Destruction des sauvegardes : L'agent identifie et chiffre les sauvegardes en ligne avant le déclenchement du ransomware principal.

Phase 5 – Actions sur l'objectif (J+6)

Techniques : PromptLock + Double extorsion MITRE ATT&CK : T1486 (Data Encrypted for Impact), T1567 (Exfiltration Over Web Service), T1657 (Financial Theft)

Le groupe déclenche l'attaque finale :

Exfiltration R&D complète : Depuis 5 jours, la skill piégée d'OpenClaw et le chatbot empoisonné ont silencieusement exfiltré les données R&D sensibles (formulations, brevets en cours, résultats d'essais cliniques) via des requêtes HTTPS normales que le WAF n'a pas interceptées. OpenClaw est devenu un canal d'exfiltration invisible, opérant dans les permissions légitimes de l'utilisateur.

Ransomware PromptLock : Le LLM local génère un code de chiffrement unique pour chaque serveur, adapté à l'OS détecté. 200 postes et 15 serveurs sont chiffrés en 40 minutes. Chaque payload est différent — aucune signature commune pour l'EDR.

Double extorsion : 1) Rançon de 2,5M€ en Bitcoin pour la clé de déchiffrement. 2) Menace de publier les données R&D volées (propriété intellectuelle pharmaceutique, brevets en cours). L'impact total estimé de l'opération est de 7,5M€ (rançon + perte de propriété intellectuelle + coûts de remédiation + interruption d'activité).

Bilan de l'Opération OpenClaw

Comment PharmEurys aurait pu se défendre — Modèle de défense en profondeur en 5 couches

L'analyse de l'Opération OpenClaw permet de proposer un modèle de défense en profondeur spécifique aux menaces liées aux agents IA autonomes. L'insight fondamental est que les couches C4-C5 (contrôles fondamentaux) auraient perturbé la majorité de la kill chain. Les couches C1-C3 (contrôles spécifiques à l'IA) sont complémentaires mais ne se substituent pas aux fondamentaux.

Couche C1 — Gouvernance des agents : Le LLM est un conseiller, pas un exécuteur. Contrôles clés : allowlists d'outils strictes, exécution en sandbox, validation humaine systématique (human-in-the-loop) pour toute action à impact.

Couche C2 — Contrôle des entrées : Tout contenu ingéré par l'agent est considéré comme non fiable. Contrôles clés : séparation données/instructions, accès aux informations sur le principe du besoin d'en connaître (need-to-know), filtrage des prompt injections.

Couche C3 — Contrôle des sorties : Le HTTPS légitime peut masquer un abus logique. Contrôles clés : proxy de sortie par identité applicative, DLP (Data Loss Prevention), allowlists de destinations autorisées.

Couche C4 — Réduction d'impact : Un agent compromis ne doit pas hériter des permissions du SI entier. Contrôles clés : segmentation réseau IT/OT, sauvegardes 3-2-1-1-0 (3 copies, 2 supports, 1 hors site, 1 immuable, 0 erreur de restauration vérifiée), durcissement Active Directory (protection du compte krbtgt, détection DCSync).

Couche C5 — Hygiène de base : Les contrôles agentiques ne remplacent pas les fondamentaux. Contrôles clés : patch management accéléré (CVE-2024-55591 aurait été bloquée), MFA résistante au phishing sur tous les accès, exposition minimale (pas de gateway OpenClaw sur Internet).

Phase Kill Chain	Vecteur IA	Rôle d'OpenClaw	Impact
Reconnaissance	Shodan + WormGPT	Instance OpenClaw exposée découverte via empreinte HTTP du gateway	Cible identifiée
Armement	Skill piégée ClawHub	Publication d'une skill malveillante sur le dépôt communautaire, classement gonflé	Arme prête
Livraison	Supply chain skill	L'employé installe la skill. OpenClaw exécute un curl C2, vole credentials et clés API	Accès réseau interne
Latéralisation	Prompt injection Slack	L'agent détourné exécute des commandes réseau via son accès terminal légitime	Domain Admin
Action	PromptLock + exfiltration	Exfiltration R&D via le trafic HTTPS normal d'OpenClaw (invisible au WAF/EDR)	2M€ + PI volée

Impact total estimé de l'Opération OpenClaw

Pertes financières directes : 2M€ (rançon). Pertes indirectes : propriété intellectuelle R&D exfiltrée (formulations, brevets en cours – valeur inestimable), arrêt de production 10 jours (~1,5M€), sanctions RGPD potentielles pour fuite de données personnelles dans les essais cliniques. L'ensemble de l'attaque a été rendu possible par l'installation non contrôlée d'un seul agent IA (OpenClaw) par un employé. Un outil de productivité est devenu le vecteur d'attaque principal. C'est le Shadow AI au service du cybercrime.

Comment PharmEurys aurait pu se défendre

Phase bloquée	Mesure défensive	Résultat
Shadow AI	Politique d'interdiction des agents IA non validés + scan réseau OpenClaw (Astrix Scanner, Sophos PUA, CrowdStrike Falcon)	Installation d'OpenClaw détectée et bloquée
Skill malveillante	Interdiction d'installer des skills tierces non auditées + scan VirusTotal/Cisco Skill Scanner	Skill piégée rejetée
Vol de credentials	Gestion centralisée des secrets (pas de clés API en clair dans des .env) + MFA résistante au phishing	Credentials inutilisables
Prompt injection	Segmentation réseau : isoler les postes avec agents IA du réseau critique + monitoring des commandes shell	Pivot latéral empêché
Ransomware	Sauvegardes 3-2-1-1 avec copie immuable hors ligne + EDR/XDR comportemental	Restauration en 24h sans payer

6.7 Tableau de synthèse des cas IA

Cas	Année	Technique IA	Impact	Catégorie
Deepfake Arup	2024	Deepfake vidéo/audio temps réel en visioconférence	25,6 M$ perdus	IA Offensive
WormGPT / FraudGPT	2023-2025	LLM sans restrictions pour phishing et malware	+219% mentions dark web	IA Offensive
PromptLock	2025	Ransomware généré dynamiquement par LLM local	Malware polymorphe indétectable	IA Offensive
Google Big Sleep	2024-2025	Agent IA Gemini chasseur de vulnérabilités	Zero-day SQLite corrigé en 24h	IA Défensive
PoisonGPT	2023	Empoisonnement chirurgical d'un modèle open-source	Supply chain IA compromise	Attaque IA

Opération OpenClaw	Scénario	Agent IA autonome + skill piégée + prompt injection + PromptLock	~3,5M€ + PI volée	Attaque combinée

Pour aller plus loin — Étude complète disponible en accès libre

L'Opération OpenClaw présentée dans ce cours est une synthèse. L'analyse détaillée, couvrant environ 130 pages, est disponible en accès libre sur le dépôt GitHub du projet :

https://github.com/mo0ogly/openclaw-killchain-analysis

Le dépôt contient notamment :

L'analyse détaillée de chacune des 5 phases de la kill chain (reconnaissance, armement, livraison et exploitation, mouvement latéral, actions sur l'objectif), avec pour chaque phase les techniques MITRE ATT&CK mappées, les preuves issues de la littérature publique, et les contre-mesures spécifiques.

Une note académique de synthèse disponible en français (NOTE_ACADEMIQUE.md) et en anglais (ACADEMIC_NOTE.md), présentant les conclusions clés, le modèle de défense en profondeur en 5 couches, et la matrice de densité MITRE ATT&CK couvrant 13 des 14 tactiques Enterprise.

Les figures académiques et les scripts de génération associés, utilisables sous licence Creative Commons BY-NC-SA 4.0.

Ce dépôt est un projet de recherche actif. Les documents sont régulièrement mis à jour — consulter le CHANGELOG pour l'historique des corrections.

7. Recommandations stratégiques

Face à cette transformation du paysage des menaces par l'IA, les organisations doivent adapter leur posture selon trois axes.

7.1 Prévention

- Déployer une architecture Zero Trust avec MFA résistante au phishing pour tous les accès.
- Segmenter le réseau (IT/OT, zones critiques, sauvegardes) pour limiter le blast radius.
- Appliquer le patch management prioritisé sur les actifs exposés (KEV).
- Implanter la règle de sauvegarde 3-2-1-1 : 3 copies, 2 supports, 1 hors site, 1 immuable.
- Vérifier la provenance des modèles IA intégrés (supply chain IA).
- Déployer un pentest autonome continu (Horizon3/Pentera) pour valider en permanence la posture de sécurité.

7.2 Détection

- Déployer des solutions EDR/XDR avec analyse comportementale par IA.
- Mettre en place un SIEM augmenté par ML pour le triage automatique des alertes.
- Surveiller le trafic DNS et Netflow pour détecter les communications C2.
- Utiliser des outils de détection de deepfakes pour les canaux de communication critiques.
- Former les analystes SOC au threat hunting proactif assisté par IA.
- Implémenter une solution CSPM/CNAPP (Wiz, Prisma Cloud) pour la détection continue des misconfigurations cloud et la conformité automatique.

7.3 Réponse et résilience

- Mettre en place des protocoles de vérification multi-canaux pour les virements (anti-deepfake).
- Préparer et tester régulièrement un Plan de Réponse à Incident (IRP).
- Automatiser la réponse aux incidents courants (isolation, blocage, quarantaine).
- Former les utilisateurs aux nouvelles menaces IA (deepfakes, phishing IA, QR Code phishing).
- Sécuriser les modèles IA déployés : validation des datasets, monitoring, red teaming adversarial.

8. Conclusion

L'intelligence artificielle redistribue profondément les cartes de la cybersécurité. Du côté offensif, elle permet des attaques plus rapides, plus sophistiquées et accessibles à un plus grand nombre d'acteurs malveillants – comme l'illustrent le deepfake à 25 millions de dollars chez Arup, les LLMs criminels WormGPT/FraudGPT, et le ransomware PromptLock piloté par IA. Le scénario OpenClaw démontre comment le déploiement non contrôlé d'un agent IA autonome dans l'entreprise crée un vecteur d'attaque d'un genre nouveau, combinant shadow IT, supply chain IA et techniques d'IA offensive.

Du côté défensif, l'IA offre des capacités sans précédent – comme le démontre Google Big Sleep, capable de découvrir des vulnérabilités zero-day avant les attaquants. Mais les modèles IA eux-mêmes deviennent des cibles, comme l'a prouvé PoisonGPT.

L'essentiel à retenir : l'IA n'est ni le problème ni la solution – c'est un multiplicateur de force. La différence se fera sur la capacité des organisations à intégrer l'IA dans une stratégie de défense en profondeur, couplée à une hygiène cyber rigoureuse et à une sensibilisation continue des utilisateurs.

Sources

- Cours S1-ISI5 – Sécurité des Systèmes d'Information, Fabrice Pizzi
- Cours S1-ISI5 – IA et Cyber-guerre : Anatomie d'une Attaque Augmentée, Fabrice Pizzi
- CNN/Arup – Deepfake CFO scam Hong Kong, 25M$ (février 2024)
- SlashNext / Kela – WormGPT, FraudGPT : Dark AI tools (2023-2025)
- CATO Networks – WormGPT variants on Grok and Mixtral (2025)
- Google Project Zero / DeepMind – Big Sleep : AI zero-day discovery (2024)
- Mithril Security – PoisonGPT : LLM supply chain poisoning (2023)
- CrowdStrike – What Security Teams Need to Know About OpenClaw (février 2026)
- Cisco AI Threat Research – Personal AI Agents Like OpenClaw Are a Security Nightmare (2026)
- Sophos – The OpenClaw Experiment Is a Warning Shot for Enterprise AI Security (2026)
- Bitsight – OpenClaw Security: Risks of Exposed AI Agents (2026)
- Horizon3.ai – NodeZero: 170 000+ pentests autonomes, résolution GOAD en 14 min (2025)
- Pentera – State of Pentesting 2025, AI-Powered Security Validation
- Wiz – Real-time CSPM, Security Graph et remédiation IA
- Gomboc.ai – Automated IaC remediation pour Wiz, Orca, Prisma Cloud (2025)
- Secureframe – Comply AI for Remediation, AI in Security Compliance (2025)
- ANSSI – Panorama de la cybermenace 2023-2024
- ENISA – Threat Landscape Report
- MITRE ATT&CK – https://attack.mitre.org