AYI NEDJIMI Consultants

IA Offensive et Défensive en Cybersécurité | Guide 2025

Catégorie : Livres Blancs Lecture : 60 min Publié le : 11/03/2026 Auteur : Ayi NEDJIMI

IA en cybersecurite : attaques adversariales, detection par ML, LLM offensifs, defense automatisee et conformite EU AI Act. Guide expert 2025.

L'intelligence artificielle transforme radicalement le paysage de la cybersécurité. D'un côté, les attaquants exploitent le machine learning pour automatiser le phishing, générer des deepfakes convaincants et créer des malwares polymorphes indétectables. De l'autre, les défenseurs déploient des modèles de détection d'anomalies, des systèmes SOAR alimentés par l'IA et des LLM spécialisés pour le threat hunting. Ce livre blanc de référence analyse en profondeur les deux faces de cette révolution technologique, depuis les attaques adversariales contre les modèles ML jusqu'aux cadres réglementaires comme l'EU AI Act et le NIST AI RMF. Destiné aux ingénieurs IA/ML, analystes sécurité, RSSI et chercheurs, il constitue un guide exhaustif pour comprendre, anticiper et contrer les menaces liées à l'intelligence artificielle en cybersécurité. Ce guide expert explore en detail les techniques d'attaque et de defense basees sur l'intelligence artificielle, les implications reglementaires du EU AI Act et les strategies de protection contre les menaces emergentes liees aux modeles generatifs.

Points clés

- L'IA offensive permet d'automatiser et de personnaliser les attaques à une échelle majeur : phishing ciblé, deepfakes, malwares génératifs et reconnaissance OSINT automatisée.
- Les attaques adversariales (evasion, poisoning, model extraction) représentent une menace critique contre les systèmes de machine learning déployés en production.
- L'IA défensive transforme la détection d'anomalies, le threat hunting et la réponse à incident grâce aux plateformes SOAR augmentées par le ML.
- Les grands modèles de langage (LLM) ouvrent de nouvelles possibilités pour l'analyse de logs, le reverse engineering et la threat intelligence automatisée.
- La sécurisation des systèmes d'IA eux-mêmes devient un enjeu majeur, encadré par l'OWASP Top 10 LLM, le MITRE ATLAS et le NIST AI RMF.
- Le cadre réglementaire européen (EU AI Act) impose des obligations strictes pour les systèmes d'IA à haut risque, y compris dans le domaine de la cybersécurité.
- Une approche équilibrée combinant IA offensive (red teaming) et défensive (blue teaming) est indispensable pour une posture de sécurité robuste.

Comment mesurez-vous concrètement l'efficacité de votre programme de sécurité ?

Chapitre 1 : Introduction - L'IA comme arme et bouclier en cybersécurité

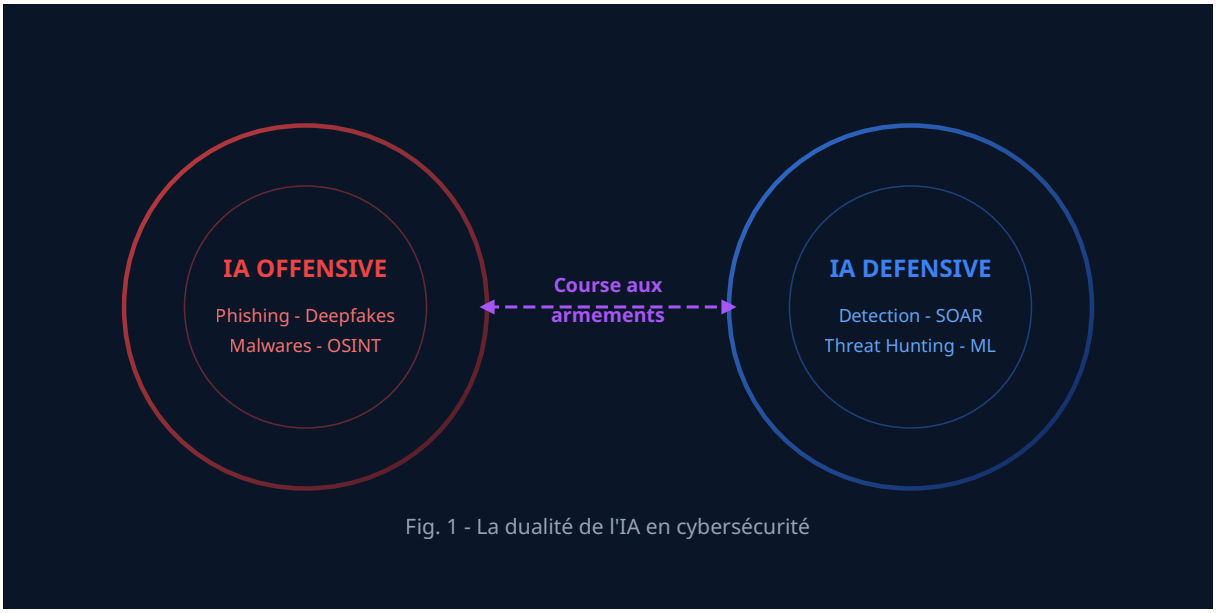

Fig. 1 - La dualité de l'IA en cybersécurité

1.1 Le cadre dual de l'intelligence artificielle

Depuis l'émergence des réseaux de neurones profonds au début des années 2010 et l'explosion des grands modèles de langage (LLM) à partir de 2022, l'intelligence artificielle a profondément reconfiguré l'écosystème de la cybersécurité. Cette transformation ne se limite pas à l'amélioration incrémentale des outils existants : elle redéfinit fondamentalement la nature même des menaces et des défenses. L'IA agit simultanément comme une arme redoutable entre les mains des attaquants et comme un bouclier poussé pour les défenseurs, créant une dynamique de course aux armements technologiques majeur dans l'histoire de la sécurité informatique.

Le rapport 2024 du World Economic Forum sur les risques globaux identifie la militarisation de l'IA comme l'une des cinq menaces technologiques majeures pour la décennie à venir. Parallèlement, le marché mondial de l'IA appliquée à la cybersécurité devrait atteindre 46,3 milliards de dollars d'ici 2027 selon MarketsandMarkets, témoignant de l'investissement massif des organisations dans les capacités défensives basées sur le machine learning. Cette dualité constitue le fil conducteur de ce livre blanc.

Définition : IA en cybersécurité
L'application de l'intelligence artificielle à la cybersécurité englobe l'ensemble des techniques de machine learning (apprentissage supervisé, non supervisé, par renforcement), de traitement du langage naturel (NLP) et de vision par ordinateur utilisées tant pour mener des cyberattaques que pour s'en défendre. Cela inclut les systèmes experts, les réseaux de neurones profonds, les modèles génératifs (GANs, LLMs) et les algorithmes de détection d'anomalies déployés dans le contexte de la sécurité des systèmes d'information.

Notre avis d'expert
Nos retours d'expérience montrent que les organisations qui investissent dans la lecture et l'application de référentiels méthodologiques structurés réduisent leur temps de réponse aux incidents de 40% en moyenne. La connaissance formalisée est un avantage compétitif sous-estimé.

1.2 L'évolution historique : du rule-based au deep learning

Pour comprendre la révolution actuelle, retracer l'évolution de l'IA en cybersécurité. Les premiers systèmes de détection d'intrusion (IDS) des années 1990, comme Snort, reposaient exclusivement sur des signatures statiques et des règles écrites manuellement. L'efficacité de ces systèmes dépendait entièrement de la capacité des analystes à anticiper les patterns d'attaque, une approche intrinsèquement réactive et limitée face à des menaces évolutives.

L'introduction du machine learning dans la cybersécurité au début des années 2000 a marqué un premier tournant. Les algorithmes de classification supervisée, notamment les arbres de décision (Random Forest) et les machines à vecteurs de support (SVM), ont permis d'automatiser la détection de malwares en analysant des caractéristiques statiques et dynamiques des fichiers exécutables. Cependant, ces approches restaient vulnérables aux techniques d'obfuscation et nécessitaient un feature engineering manuel considérable.

La véritable rupture est intervenue avec l'adoption du deep learning à partir de 2015-2016. Les réseaux de neurones convolutifs (CNN) et récurrents (LSTM, GRU) ont démontré leur capacité à extraire automatiquement des caractéristiques pertinentes à partir de données brutes, qu'il s'agisse de flux réseau, de séquences d'appels système ou de code binaire. Des travaux pionniers comme ceux de Raff et al. (2018) sur la détection de malwares à partir de fichiers PE bruts avec des réseaux convolutifs profonds ont établi de nouveaux standards de performance, dépassant les approches traditionnelles basées sur des signatures.

L'émergence des modèles de langage pré-entraînés, culminant avec GPT-4, Claude et les modèles open source comme Llama et Mistral, a ouvert une nouvelle ère. Ces modèles démontrent des capacités remarquables pour l'analyse de code, la compréhension de vulnérabilités, la génération de rapports de threat intelligence et même l'assistance au reverse engineering. Simultanément, ils offrent aux attaquants des outils puissants pour automatiser la création de contenu de phishing persuasif, générer du code malveillant et conduire des opérations de manipulation informationnelle à grande échelle.

Votre stratégie de cybersécurité repose-t-elle sur un référentiel méthodologique éprouvé ?

1.3 Cartographie des acteurs et des menaces

L'écosystème des menaces liées à l'IA en cybersécurité implique une diversité d'acteurs aux motivations et capacités variées. Les groupes APT (Advanced Persistent Threat) étatiques, disposant de ressources considérables, intègrent progressivement des capacités d'IA dans leurs arsenaux offensifs. Le rapport Mandiant 2024 documente l'utilisation par des groupes affiliés à la Chine (APT41), à la Russie (APT28/Fancy Bear) et à la Corée du Nord (Lazarus Group) de techniques assistées par l'IA pour la reconnaissance, le spear-phishing et l'évasion de détection.

Les organisations cybercriminelles adoptent également l'IA à des fins lucratives. L'apparition d'outils comme WormGPT et FraudGPT sur les forums du dark web illustre la démocratisation des capacités offensives basées sur l'IA. Ces outils, dérivés de modèles de langage open source fine-tunés sur des données malveillantes, permettent à des attaquants peu qualifiés de générer des emails de phishing avancés, du code d'exploitation et des scripts d'attaque personnalisés.

Contexte : Le marché de l'IA offensive sur le dark web
Depuis mi-2023, plusieurs services d'IA offensive sont apparus sur les marchés souterrains. WormGPT, basé sur GPT-J 6B fine-tuné, propose la génération d'emails de phishing et de code malveillant pour un abonnement mensuel d'environ 60 euros. FraudGPT, similaire dans son fonctionnement, cible spécifiquement la fraude financière et le carding. DarkBART et DarkBERT représentent des adaptations de modèles existants entraînés sur des données du dark web. Ces outils, bien que souvent surestimés dans leurs capacités réelles, symbolisent une tendance de fond vers la commoditisation de l'IA offensive.

Cas concret
L'ANSSI a publié en 2023 son guide de recommandations pour l'administration sécurisée des SI, mettant à jour les principes de Tiering et de bastionnement. Ce document de référence pour les organisations françaises rappelle que les fondamentaux de l'hygiène informatique restent les mesures les plus efficaces.

1.4 Les enjeux stratégiques pour les organisations

Face à cette dualité, les organisations sont confrontées à un triple défi. Premièrement, elles doivent intégrer l'IA dans leurs capacités défensives pour maintenir leur posture de sécurité face à des menaces de plus en plus élaborées. Deuxièmement, elles doivent anticiper et se préparer aux attaques utilisant l'IA, ce qui implique de comprendre en profondeur les techniques offensives. Troisièmement, elles doivent sécuriser leurs propres déploiements d'IA contre les attaques adversariales, le data poisoning et les fuites de données.

Ce livre blanc aborde systématiquement ces trois dimensions. Les chapitres 2 à 4 explorent les capacités offensives de l'IA. Les chapitres 5 à 7 détaillent les applications défensives. Le chapitre 8 traite de la sécurisation des systèmes d'IA eux-mêmes. Enfin, le chapitre 9 analyse le cadre éthique et réglementaire, notamment l'EU AI Act et le NIST AI RMF, qui encadrent l'utilisation de l'IA en cybersécurité.

A retenir - Chapitre 1
L'IA en cybersécurité est fondamentalement duale : elle amplifie simultanément les capacités offensives et défensives. La course aux armements entre attaquants et défenseurs s'accélère avec chaque avancée technologique, de la détection d'anomalies par deep learning aux LLM génératifs. Les organisations doivent adopter une approche holistique intégrant l'IA dans leur défense, anticipant les attaques basées sur l'IA et sécurisant leurs propres systèmes d'IA.

Chapitre 2 : IA offensive - Phishing automatisé, deepfakes et génération de malwares

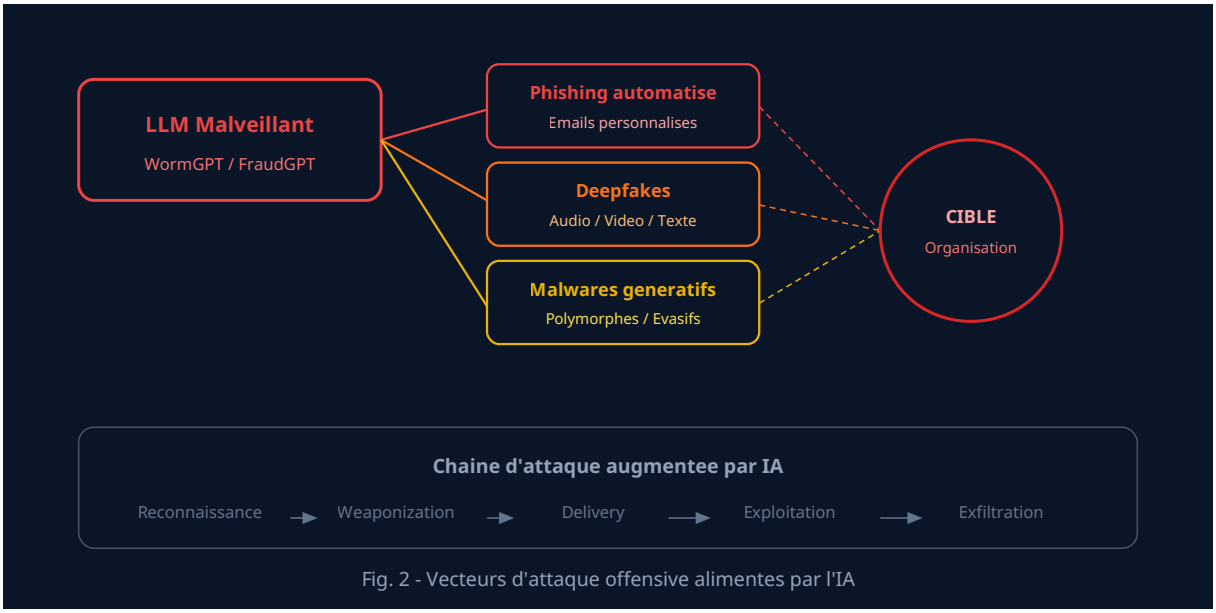

Fig. 2 - Vecteurs d'attaque offensive alimentes par l'IA

2.1 Phishing automatisé et ingénierie sociale augmentée par l'IA

Le phishing demeure le vecteur d'attaque initial le plus répandu, représentant selon le rapport Verizon DBIR 2024 plus de 36% des compromissions initiales. L'intégration de l'IA dans les campagnes de phishing marque une rupture qualitative majeure par rapport aux approches traditionnelles. Alors que le phishing classique reposait sur des templates génériques facilement identifiables par les utilisateurs avertis et les filtres anti-spam, le phishing augmenté par l'IA permet une personnalisation à grande échelle, rendant chaque message unique et contextuellement pertinent pour sa cible.

Les grands modèles de langage permettent de générer des emails de spear-phishing d'une qualité linguistique irréprochable, adaptés au contexte professionnel, au style de communication et aux centres d'intérêt de chaque cible. Une étude menée par des chercheurs de l'Université de Singapour (Heiding et al., 2023) a démontré que les emails de phishing générés par GPT-4 obtenaient un taux de clic supérieur de 30% par rapport aux emails rédigés par des attaquants humains expérimentés. Plus inquiétant encore, ces emails étaient moins fréquemment détectés par les solutions de sécurité traditionnelles, les modèles génératifs produisant des variations textuelles suffisamment diversifiées pour contourner les filtres basés sur des patterns connus.

La chaîne d'attaque du phishing augmenté par l'IA se décompose en plusieurs étapes automatisées. Premièrement, la phase de reconnaissance utilise des techniques d'OSINT (Open Source Intelligence) automatisées pour collecter des informations sur la cible : profils LinkedIn, publications sur les réseaux sociaux, organigrammes d'entreprise, communiqués de presse. Deuxièmement, le LLM génère un email personnalisé intégrant ces éléments contextuels,

adoptant le ton et le style appropriés. Troisièmement, le système adapte automatiquement le prétexte en fonction du rôle de la cible : facture urgente pour un comptable, mise à jour de sécurité pour un administrateur système, invitation à une conférence pour un dirigeant.

Alerte : Le vishing augmenté par l'IA
Au-delà du phishing par email, l'IA permet désormais le vishing (voice phishing) automatisé grâce à la synthèse vocale en temps réel. Des modèles comme VALL-E (Microsoft Research) et XTTS (Coqui) peuvent cloner une voix à partir de quelques secondes d'échantillon audio. En 2024, une entreprise de Hong Kong a été victime d'une fraude de 25 millions de dollars après qu'un employé a participé à une visioconférence où tous les autres participants, y compris le CFO, étaient des deepfakes en temps réel. Cette attaque illustre la convergence entre deepfakes vidéo, synthèse vocale et ingénierie sociale automatisée.

2.2 Deepfakes : la manipulation de la réalité numérique

Les deepfakes, produits par des réseaux antagonistes génératifs (GANs) et des modèles de diffusion, représentent une menace croissante en cybersécurité. Ces contenus synthétiques, qu'ils soient visuels, audio ou textuels, permettent l'usurpation d'identité à une échelle et avec un réalisme majeur. La technologie de deepfake a considérablement progressé depuis les premiers travaux de Goodfellow et al. (2014) sur les GANs, atteignant un niveau de réalisme qui rend la détection visuelle quasi impossible pour un observateur humain non entraîné.

Dans le contexte de la cybersécurité offensive, les deepfakes sont utilisés selon plusieurs vecteurs. Le premier concerne la fraude au dirigeant (CEO fraud) par deepfake audio ou vidéo. Les attaquants synthétisent la voix du PDG ou du directeur financier pour ordonner des virements frauduleux ou communiquer de fausses instructions à des collaborateurs. Le second vecteur concerne la désinformation et la manipulation d'opinion, utilisées dans le cadre d'opérations d'influence étatiques ou de campagnes de déstabilisation d'entreprises. Le troisième vecteur est le contournement des systèmes d'authentification biométrique basés sur la reconnaissance faciale ou vocale.

Les modèles de diffusion de dernière génération, comme Stable Diffusion XL et DALL-E 3, permettent de créer des images photoréalistes de personnes fictives ou de manipuler des images existantes avec une précision remarquable. Les outils de face-swap en temps réel, comme DeepFaceLive, permettent de conduire des appels vidéo en se faisant passer pour une autre personne. Les implications pour la sécurité sont considérables : les processus de vérification d'identité par visioconférence, les systèmes KYC (Know Your Customer) des institutions financières et les mécanismes d'authentification multi-facteurs basés sur la biométrie faciale deviennent tous potentiellement vulnérables.

Type de deepfake	Technologie	Vecteur d'attaque	Détection
Audio (voice cloning)	VALL-E, XTTS, RVC	Vishing, fraude au dirigeant	Analyse spectrale, watermarking
Vidéo (face swap)	DeepFaceLab, FaceSwap	Usurpation visioconférence	Détection de micro-expressions, artefacts
Vidéo temps réel	DeepFaceLive, Avatarify	KYC bypass, ingénierie sociale	Challenge liveness, 3D depth
Image	Stable Diffusion, DALL-E 3	Faux documents, identités fictives	Analyse métadonnées, détection GAN
Texte	GPT-4, Claude, Llama	Phishing, désinformation	Détection IA (GPTZero, Originality)

2.3 Génération de malwares par IA : WormGPT, FraudGPT et au-delà

L'utilisation de l'IA pour la génération de code malveillant représente une évolution majeure dans l'écosystème des menaces. Si les modèles de langage commerciaux comme GPT-4 et Claude intègrent des garde-fous (guardrails) destinés à empêcher la génération de contenu malveillant, ces protections peuvent être contournées par des techniques de jailbreaking ou tout simplement évitées en utilisant des modèles open source sans restrictions éthiques. L'écosystème criminel a rapidement développé des alternatives dédiées.

`WormGPT`, apparu en juillet 2023, constitue le premier exemple médiatisé d'un LLM spécifiquement conçu pour la cybercriminalité. Basé sur le modèle GPT-J 6B de EleutherAI, fine-tuné sur des données relatives aux malwares et aux techniques d'attaque, WormGPT permet de générer des emails de phishing convaincants, des scripts d'exploitation et des payloads malveillants sans les restrictions éthiques des modèles commerciaux. `FraudGPT`, apparu peu après, cible spécifiquement les activités de fraude financière : génération de pages de phishing imitant des interfaces bancaires, création de scripts de carding et rédaction de messages d'arnaque.

Au-delà de ces outils médiatisés, la menace réelle réside dans la capacité des modèles de langage à assister le développement de malwares polymorphes. Le concept de malware polymorphe assisté par IA repose sur l'utilisation d'un LLM pour réécrire automatiquement le code malveillant à chaque itération, modifiant la structure syntaxique et les signatures binaires tout en préservant la fonctionnalité malveillante. Des chercheurs de HYAS Labs ont démontré en 2023 avec leur proof-of-concept BlackMamba qu'un keylogger pouvait utiliser l'API de GPT pour modifier dynamiquement son code à chaque exécution, rendant la détection par signatures quasi impossible.

Les techniques de génération de malwares par IA incluent également l'obfuscation automatique de code existant, la génération de techniques d'évasion de sandbox, la création de communications C2 (Command and Control) imitant du trafic légitime, et la synthèse de chaînes

d'exploitation (exploit chains) combinant plusieurs vulnérabilités. Le framework MITRE ATLAS (Adversarial Threat Landscape for AI Systems) documente systématiquement ces techniques dans sa matrice de tactiques et procédures adversariales spécifiques à l'IA.

"Les modèles de langage actuels peuvent générer du code fonctionnel qui, avec un minimum d'adaptation humaine, peut être transformé en outil offensif. La barrière à l'entrée pour la cybercriminalité s'en trouve significativement abaissée, même si les capacités réelles de ces outils sont souvent surévaluées par rapport au battage médiatique."

-- Rapport EUROPOL, Chatbots and Criminal Use of Large Language Models, 2024

2.4 Automatisation de la kill chain par l'IA

L'IA ne se contente pas d'améliorer individuellement chaque phase d'une cyberattaque : elle permet d'automatiser l'intégralité de la kill chain, de la reconnaissance initiale à l'exfiltration de données. Le concept d'attaque autonome (autonomous cyber attack) fait l'objet de recherches tant académiques que militaires, et soulève des questions fondamentales sur la nature future des conflits cyber.

Dans la phase de reconnaissance, l'IA automatise la collecte et l'analyse d'informations sur la cible. Des outils combinant web scraping, analyse de réseaux sociaux et corrélation de données publiques permettent de construire automatiquement un profil détaillé de l'infrastructure technique et de l'organigramme d'une organisation. Dans la phase de weaponization, les LLM génèrent des payloads adaptés aux vulnérabilités identifiées. Dans la phase de delivery, l'IA optimise le timing, le canal et le contenu des vecteurs d'attaque pour maximiser la probabilité de compromission. Dans les phases d'exploitation et de post-exploitation, des agents IA peuvent théoriquement pivoter latéralement dans un réseau, identifier les actifs de valeur et exfiltrer les données de manière autonome.

Le projet AutoAttacker, présenté par des chercheurs de l'Université de l'Illinois en 2024, a démontré qu'un agent basé sur GPT-4 pouvait exploiter automatiquement des vulnérabilités connues dans des environnements de test, réalisant des chaînes d'exploitation multi-étapes sans intervention humaine. Bien que limité à des scénarios contrôlés, ce travail illustre le potentiel des agents IA autonomes dans un contexte offensif. De même, le benchmark CyberBench évalue les capacités des LLM sur des tâches offensives réelles, révélant que les modèles les plus avancés peuvent résoudre des challenges CTF (Capture The Flag) de niveau intermédiaire de manière autonome.

A retenir - Chapitre 2
L'IA offensive amplifie considérablement les capacités des attaquants à travers trois vecteurs principaux : le phishing personnalisé à grande échelle, les deepfakes pour l'usurpation d'identité et la génération automatisée de malwares polymorphes. Des outils comme WormGPT et FraudGPT démocratisent l'accès à ces capacités. L'automatisation complète de la kill chain par des agents IA autonomes, bien qu'encore expérimentale, représente une menace émergente que les défenseurs doivent anticiper.

Chapitre 3 : Attaques adversariales contre les modèles de machine learning

Fig. 3 - Taxonomie des attaques adversariales contre les modeles ML

3.1 Evasion attacks : tromper les modèles en production

Les attaques par évasion (evasion attacks) constituent la catégorie la plus étudiée et la plus immédiatement applicable d'attaques adversariales. Leur principe consiste à modifier subtilement les données d'entrée d'un modèle de machine learning déployé en production afin de provoquer une classification erronée, tout en maintenant les modifications imperceptibles pour un observateur humain. Ces perturbations adversariales exploitent les propriétés géométriques des frontières de décision apprises par les réseaux de neurones profonds.

Les travaux fondateurs de Szegedy et al. (2013) ont révélé que l'ajout de perturbations quasi imperceptibles aux images d'entrée pouvait faire basculer la prédiction d'un réseau de neurones profond avec une confiance élevée. Goodfellow et al. (2014) ont formalisé cette observation avec la méthode FGSM (Fast Gradient Sign Method), qui calcule la perturbation optimale en un seul pas de gradient. Des méthodes plus complexes ont ensuite été développées : PGD (Projected Gradient Descent) de Madry et al. (2018), les attaques C&W (Carlini and Wagner, 2017), et DeepFool (Moosavi-Dezfooli et al., 2016), chacune offrant différents compromis entre efficacité de l'attaque, imperceptibilité de la perturbation et coût computationnel.

Dans le contexte de la cybersécurité, les attaques par évasion ont des implications directes et critiques. Un malware peut être modifié pour échapper à un classificateur basé sur le deep learning en ajoutant des octets de padding calculés de manière adversariale, sans altérer sa fonctionnalité malveillante. Les travaux de Kolosnjaji et al. (2018) et Anderson et al. (2018) ont démontré la faisabilité de ces attaques contre des détecteurs de malwares basés sur des réseaux de neurones, avec des taux d'évasion supérieurs à 60% sur des modèles de production. Le framework `MalwareGym`, développé par des chercheurs de CrowdStrike, formalise cette

approche en proposant un environnement d'apprentissage par renforcement où un agent IA apprend automatiquement les modifications minimales nécessaires pour contourner un détecteur donné.

La propriété de transférabilité des exemples adversariaux aggrave considérablement cette menace. Un exemple adversarial conçu pour tromper un modèle A peut souvent tromper également un modèle B, même si les deux modèles ont des architectures différentes et ont été entraînés sur des données distinctes. Cette propriété, démontrée par Papernot et al. (2016), permet des attaques en boîte noire (black-box attacks) où l'attaquant n'a pas besoin de connaître les détails du modèle cible. Il lui suffit d'entraîner un modèle substitut et de générer des exemples adversariaux sur ce substitut, qui se transféreront au modèle cible avec une probabilité significative.

Exemple concret : Adversarial patches physiques
Les adversarial patches démontrent que les attaques adversariales ne se limitent pas au domaine numérique. Des chercheurs de l'Université de Leuven ont montré qu'un patch imprimé sur un t-shirt pouvait rendre une personne invisible pour les détecteurs d'objets basés sur YOLO. Dans un contexte de sécurité physique, cette technique pourrait être utilisée pour contourner des systèmes de vidéosurveillance intelligente. Eykholt et al. (2018) ont démontré que de simples stickers placés sur des panneaux de signalisation pouvaient tromper les systèmes de conduite autonome, faisant classifier un panneau stop comme une limitation de vitesse (CVE relaté dans le cadre des travaux sur la sécurité des véhicules autonomes).

3.2 Data poisoning : corrompre les données d'entraînement

Les attaques par empoisonnement de données (data poisoning) ciblent la phase d'entraînement des modèles de machine learning plutôt que leur phase d'inférence. En injectant des données malveillantes dans le jeu d'entraînement, un attaquant peut compromettre le comportement du modèle résultant, soit en dégradant ses performances globales, soit en introduisant une porte dérobée (backdoor) qui sera activée par un déclencheur spécifique lors de l'inférence.

Les attaques par backdoor sont particulièrement insidieuses. Le modèle empoisonné fonctionne normalement sur les données légitimes, rendant sa compromission difficile à détecter par les métriques de performance standard. Cependant, lorsqu'un pattern de déclenchement spécifique (trigger pattern) est présent dans les données d'entrée, le modèle produit la sortie souhaitée par l'attaquant. Gu et al. (2017) ont introduit le concept de BadNets, démontrant qu'un réseau de neurones pouvait être entraîné avec une backdoor activée par un petit motif visuel ajouté aux images d'entrée. Les clean-label attacks, une variante plus aboutie introduite par Shafahi et al. (2018), permettent d'empoisonner un modèle sans même modifier les labels des données d'entraînement, rendant la détection encore plus complexe.

Dans le contexte de la cybersécurité, le data poisoning menace directement les systèmes de détection d'intrusion et d'anti-malware basés sur le ML. Si un attaquant parvient à influencer les données d'entraînement d'un système de détection d'anomalies réseau, par exemple en injectant progressivement du trafic malveillant dans les données considérées comme normales, il peut entraîner le système à considérer son trafic d'attaque comme légitime. Les modèles

entraînés sur des données issues de sources ouvertes (threat intelligence feeds, bases de signatures communautaires) sont particulièrement vulnérables si l'intégrité de ces sources n'est pas rigoureusement vérifiée.

Le poisoning des modèles de langage (LLM poisoning) représente une variante émergente de cette menace. Des chercheurs ont démontré qu'il était possible d'injecter des backdoors dans des LLM pendant la phase de fine-tuning, créant des modèles qui génèrent du code vulnérable ou divulguent des informations sensibles en réponse à des prompts spécifiques. Le risque est amplifié par la pratique courante du fine-tuning de modèles pré-entraînés sur des données potentiellement non vérifiées, ainsi que par l'utilisation de modèles open source provenant de sources tierces dont l'intégrité ne peut être garantie.

3.3 Model extraction et model inversion

Les attaques par extraction de modèle (model extraction) visent à voler la propriété intellectuelle représentée par un modèle de ML déployé en tant que service (MLaaS). L'attaquant, n'ayant accès qu'à l'API du modèle cible, soumet un ensemble de requêtes soigneusement conçues et utilise les réponses pour entraîner un modèle substitut qui reproduit le comportement du modèle original. Tramer et al. (2016) ont démontré la faisabilité de cette approche contre des modèles déployés sur Amazon ML, BigML et Google Prediction API, répliquant des modèles de régression logistique et d'arbres de décision avec une fidélité quasi parfaite en quelques milliers de requêtes.

Pour les modèles plus complexes comme les réseaux de neurones profonds, l'extraction nécessite davantage de requêtes mais reste réalisable. Krishna et al. (2020) ont démontré l'extraction de modèles BERT fine-tunés avec une fidélité supérieure à 95% en utilisant des techniques de distillation de connaissances adaptées. Orekondy et al. (2019) ont introduit Knockoff Nets, une méthode d'extraction de modèles de vision par ordinateur utilisant des données de substitution sans rapport avec les données d'entraînement originales, démontrant que même sans connaissance du domaine, un attaquant peut reproduire efficacement un modèle cible.

Les attaques par inversion de modèle (model inversion) constituent une menace pour la confidentialité des données d'entraînement. Fredrikson et al. (2015) ont démontré qu'il était possible de reconstruire des images de visages à partir d'un modèle de reconnaissance faciale, simplement en interrogeant le modèle. Les membership inference attacks, introduites par Shokri et al. (2017), permettent de déterminer si un échantillon spécifique faisait partie des données d'entraînement, posant des risques significatifs pour la protection des données personnelles, notamment dans les contextes soumis au RGPD.

Type d'attaque	Phase ciblée	Objectif	Défense principale	Référence
FGSM / PGD	Inférence	Evasion de classification	Adversarial training	Goodfellow et al., 2014 / Madry et al., 2018
C&W Attack	Inférence	Evasion avec perturbation minimale	Distillation défensive	Carlini & Wagner, 2017
BadNets	Entraînement	Backdoor par trigger pattern	Neural Cleanse, pruning	Gu et al., 2017
Clean-label poisoning	Entraînement	Backdoor sans modification de labels	Spectral signatures	Shafahi et al., 2018
Model extraction	Inférence (API)	Vol de propriété intellectuelle	Watermarking, rate limiting	Tramer et al., 2016
Model inversion	Inférence (API)	Reconstruction de données privées	Differential privacy	Fredrikson et al., 2015
Membership inference	Inférence (API)	Audit d'appartenance aux données	Regularization, DP-SGD	Shokri et al., 2017

3.4 Défenses contre les attaques adversariales

La recherche en robustesse adversariale a produit un arsenal de techniques défensives, bien qu'aucune ne constitue une solution universelle. L'adversarial training, proposé par Madry et al. (2018), consiste à inclure des exemples adversariaux dans les données d'entraînement afin de rendre le modèle robuste à ces perturbations. Cette approche, bien qu'efficace, augmente significativement le coût d'entraînement et peut dégrader les performances sur les données non perturbées (robustness-accuracy trade-off). La distillation défensive, proposée par Papernot et al. (2016), utilise un processus d'entraînement en deux étapes pour produire des modèles plus résistants, bien que des attaques ultérieures aient démontré ses limitations.

Pour les attaques par empoisonnement, les défenses reposent principalement sur l'inspection et le nettoyage des données d'entraînement. Des techniques comme Neural Cleanse (Wang et al., 2019) et Activation Clustering (Chen et al., 2018) permettent de détecter et de supprimer les backdoors dans les modèles déjà entraînés. Le pruning neuronal, qui consiste à supprimer les neurones dormants potentiellement liés à des backdoors, offre une approche complémentaire. Contre l'extraction de modèle, les défenses incluent le watermarking des modèles, la détection de requêtes suspectes par analyse de distribution, et les techniques de differential privacy (DP-SGD) pour protéger la confidentialité des données d'entraînement.

Le framework MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems), inspiré du célèbre MITRE ATT&CK, propose une taxonomie complète des tactiques, techniques et procédures (TTPs) adversariales spécifiques aux systèmes d'IA. Ce cadre de référence, régulièrement mis à jour, permet aux organisations d'évaluer leur exposition aux attaques

adversariales et de planifier les mesures de mitigation appropriées. Il constitue un outil indispensable pour les équipes de sécurité intégrant des composants de ML dans leur infrastructure.

A retenir - Chapitre 3
Les attaques adversariales menacent les modèles ML à chaque phase de leur cycle de vie : évasion en inférence, empoisonnement en entraînement, extraction et inversion via les API. La transférabilité des exemples adversariaux et la sophistication croissante des techniques de poisoning rendent ces menaces particulièrement préoccupantes pour les systèmes de cybersécurité basés sur le ML. Le framework MITRE ATLAS fournit une taxonomie de référence pour évaluer et atténuer ces risques.

Chapitre 4 : IA pour la reconnaissance et l'OSINT automatisé

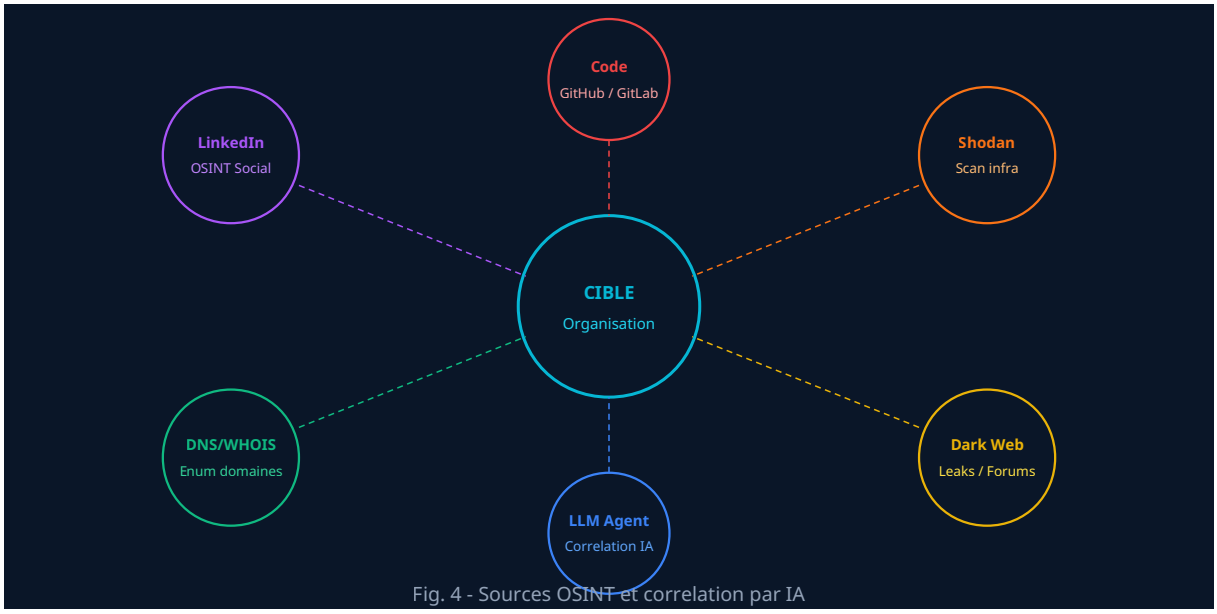

Fig. 4 - Sources OSINT et correlation par IA

4.1 La reconnaissance automatisée : de la collecte à l'analyse

La phase de reconnaissance constitue la première étape de toute opération offensive et conditionne la réussite des phases ultérieures. Traditionnellement réalisée manuellement par des pentesters et des opérateurs offensifs, cette phase est profondément transformée par l'intégration de l'intelligence artificielle. L'IA ne se contente pas d'accélérer la collecte d'informations : elle apporte une capacité d'analyse, de corrélation et de contextualisation qui permet d'extraire des insights exploitables à partir de volumes massifs de données hétérogènes.

L'OSINT (Open Source Intelligence) automatisé par IA combine plusieurs disciplines : le web scraping intelligent, le traitement du langage naturel pour l'analyse de contenu textuel, la vision par ordinateur pour l'analyse d'images et de documents, et les techniques de graph analytics pour la cartographie des relations entre entités. Des frameworks comme `SpiderFoot`, `Maltego` et `theHarvester` intègrent progressivement des capacités de ML pour automatiser la corrélation d'informations provenant de sources multiples.

Dans le domaine de la reconnaissance technique, l'IA permet d'analyser automatiquement les résultats de scans réseau (`Nmap`), d'identifier les technologies web utilisées par une cible (`Wappalyzer`), de détecter les certificats SSL et leurs chaînes de confiance, et de corréler ces informations avec les bases de vulnérabilités connues (NVD, CVE). Des agents basés sur des LLM peuvent désormais interpréter les résultats de ces outils, identifier les vecteurs d'attaque potentiels et prioriser les cibles en fonction de leur exposition et de leur criticité estimée.

4.2 OSINT sur les réseaux sociaux et profilage par IA

Les réseaux sociaux constituent une source d'information inépuisable pour la reconnaissance offensive. LinkedIn, en particulier, fournit des données précieuses sur la structure organisationnelle d'une cible : organigramme, technologies utilisées (mentionnées dans les offres d'emploi et les profils techniques), partenaires et fournisseurs, et informations sur les processus internes. L'IA permet d'automatiser l'extraction et l'analyse de ces informations à grande échelle.

Les techniques de profilage par IA vont au-delà de la simple collecte de données factuelles. Des modèles de NLP analysent le style d'écriture, les centres d'intérêt et les interactions en ligne d'une cible pour construire un profil psychologique exploitable dans le cadre d'opérations d'ingénierie sociale. Les travaux de Kosinski et al. (2013) ont démontré que l'analyse des "likes" Facebook permettait de prédire avec précision des traits de personnalité, l'orientation politique et d'autres caractéristiques personnelles, des informations directement exploitables pour la conception de campagnes de phishing ciblées.

L'analyse de graphes sociaux par des algorithmes de ML permet d'identifier les individus occupant des positions clés dans le réseau relationnel d'une organisation, les relations informelles entre départements, et les points de vulnérabilité humaine dans la chaîne de sécurité. Des techniques de community detection et de centrality analysis identifient les personnes les plus influentes ou les plus connectées, qui constituent des cibles privilégiées pour le spear-phishing et les attaques par watering hole.

Risque : Reconnaissance automatisée des dépôts de code
Les dépôts de code publics (GitHub, GitLab, Bitbucket) constituent une mine d'or pour la reconnaissance offensive. Des outils comme `TruffleHog`, `GitLeaks` et `git-secrets` détectent automatiquement les clés API, tokens d'authentification, mots de passe et certificats accidentellement commités dans les dépôts publics. L'IA augmente ces outils en analysant le contexte du code pour identifier des informations sensibles qui ne correspondent pas aux patterns de secrets classiques : noms de serveurs internes, schémas de bases de données, commentaires révélant des vulnérabilités connues ou des contournements de sécurité temporaires jamais corrigés.

4.3 Scan de vulnérabilités augmenté par IA

L'analyse automatisée de vulnérabilités bénéficie considérablement de l'intégration du ML. Les scanners de vulnérabilités traditionnels comme `Nessus`, `OpenVAS` et `Qualys` reposent principalement sur des bases de signatures et des heuristiques prédéfinies. L'IA apporte une

couche d'intelligence supplémentaire en permettant la découverte de vulnérabilités inconnues (zero-day) par analyse de patterns, la priorisation des vulnérabilités en fonction du contexte spécifique de la cible, et la génération automatique de scénarios d'exploitation.

Les modèles de NLP appliqués à l'analyse de code source permettent d'identifier des patterns de vulnérabilités qui échappent aux analyses statiques traditionnelles. Des travaux comme ceux de Li et al. (2018) avec VulDeePecker et de Zhou et al. (2019) avec Devign démontrent que les réseaux de neurones peuvent apprendre à détecter des vulnérabilités à partir du code source avec des taux de détection significativement supérieurs aux outils d'analyse statique classiques. Ces approches sont particulièrement efficaces pour les classes de vulnérabilités impliquant des interactions complexes entre plusieurs fonctions ou modules, comme les use-after-free, les race conditions et les dépassements de tampons dépendant du contexte.

L'analyse de surface d'attaque automatisée par IA intègre la reconnaissance réseau, l'identification des services exposés, l'analyse des configurations et la corrélation avec les bases de vulnérabilités pour produire une cartographie dynamique de la surface d'attaque d'une organisation. Des plateformes comme Censys, Shodan et Binary Edge fournissent les données brutes, tandis que les modèles de ML assurent l'analyse, la priorisation et la contextualisation. Le concept d'Attack Surface Management (ASM) basé sur l'IA est devenu une catégorie de produits à part entière dans l'écosystème de la cybersécurité.

4.4 Agents IA autonomes pour le pentesting

L'émergence d'agents IA autonomes capables de conduire des tests d'intrusion de manière semi-automatisée représente une avancée significative. Des projets de recherche comme PentestGPT (Deng et al., 2023) et AutoPT explorent l'utilisation de LLM comme moteur de raisonnement pour guider des opérations de pentesting. Ces agents peuvent interpréter les résultats d'outils de reconnaissance, formuler des hypothèses d'attaque, sélectionner et exécuter les outils appropriés, et adapter leur stratégie en fonction des résultats obtenus.

Le framework `PentestGPT` décompose le processus de pentesting en trois modules : un module de raisonnement basé sur GPT-4 qui planifie la stratégie d'attaque, un module de génération qui produit les commandes et scripts nécessaires, et un module de parsing qui interprète les résultats. Des évaluations sur des machines HackTheBox et des environnements CTF montrent que ces agents peuvent résoudre des scénarios de complexité intermédiaire de manière largement autonome, bien qu'ils nécessitent encore une supervision humaine pour les scénarios complexes impliquant du raisonnement créatif ou des techniques non conventionnelles.

Les implications pour la sécurité sont ambivalentes. D'un côté, ces agents démocratisent l'accès aux capacités de pentesting, permettant aux organisations disposant de ressources limitées de conduire des évaluations de sécurité plus fréquentes et plus exhaustives. De l'autre, ils abaissent la barrière d'entrée pour les acteurs malveillants, permettant à des attaquants moins qualifiés de mener des opérations offensives poussées. La communauté de la cybersécurité doit anticiper cette dualité et développer des défenses adaptées aux attaques semi-automatisées conduites par des agents IA.

A retenir - Chapitre 4

L'IA transforme la reconnaissance offensive en automatisant la collecte, la corrélation et l'analyse de données OSINT à grande échelle. L'analyse des réseaux sociaux, des dépôts de code publics et des surfaces d'attaque bénéficie directement du ML. Les agents IA autonomes pour le pentesting, bien qu'encore limités aux scénarios de complexité intermédiaire, annoncent une démocratisation des capacités offensives qui requiert une adaptation des stratégies défensives.

Chapitre 5 : IA défensive - Détection d'anomalies et threat hunting assisté par ML

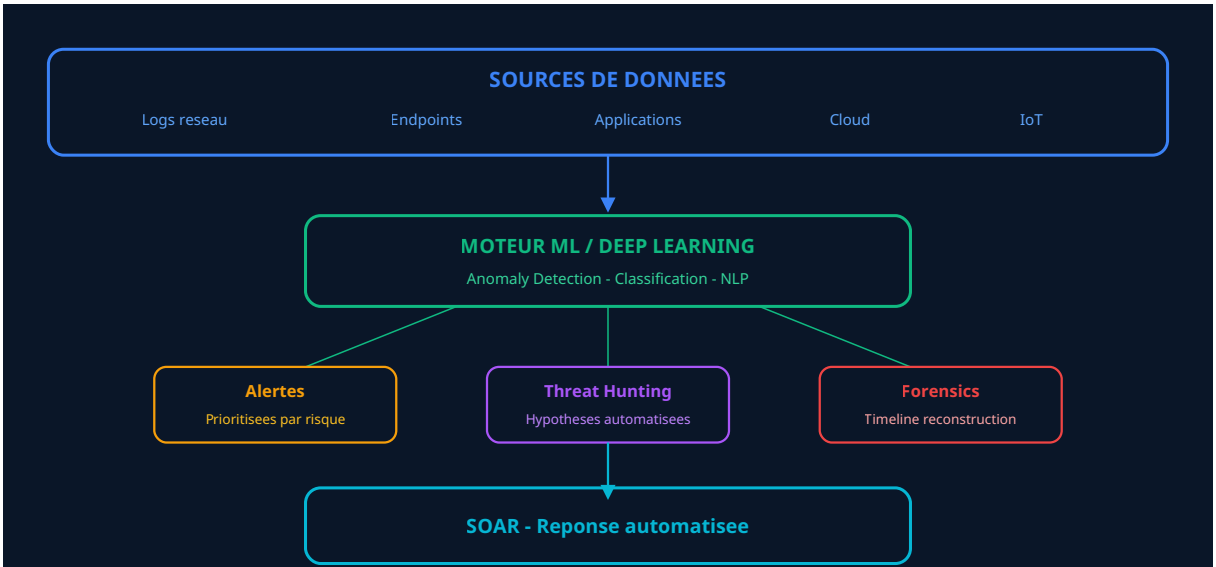

Fig. 5 - Pipeline de defense IA : de la collecte a la reponse

5.1 Détection d'anomalies réseau par machine learning

La détection d'anomalies réseau constitue l'application la plus mature et la plus répandue de l'IA défensive en cybersécurité. Contrairement aux systèmes de détection basés sur des signatures, qui ne peuvent identifier que des menaces préalablement répertoriées, les systèmes basés sur le ML apprennent le comportement normal du réseau et détectent les déviations statistiquement significatives, permettant théoriquement la détection de menaces inconnues (zero-day) et d'attaques avancées conçues pour contourner les signatures existantes.

Les approches non supervisées dominent ce domaine, car elles ne nécessitent pas de données étiquetées d'attaques, dont l'obtention est coûteuse et dont la représentativité est limitée. Les autoencodeurs, et plus particulièrement les autoencodeurs variationnels (VAE), sont largement utilisés pour apprendre une représentation compressée du trafic réseau normal. Les anomalies sont détectées par un score de reconstruction élevé, indiquant que le trafic observé dévie significativement du modèle appris. Les travaux de Mirsky et al. (2018) avec Kitsune ont démontré l'efficacité d'un ensemble d'autoencodeurs entraînés de manière incrémentale pour la détection d'anomalies réseau en temps réel, sans nécessiter de phase d'apprentissage supervisé.

Les modèles basés sur les graphes offrent une perspective complémentaire en modélisant les interactions entre entités du réseau (machines, utilisateurs, services) plutôt que les flux individuels. Les Graph Neural Networks (GNNs) et les techniques de graph embedding permettent de détecter des patterns d'attaque distribués qui seraient invisibles dans l'analyse de flux individuels, comme les mouvements latéraux, les communications C2 low-and-slow, et les exfiltrations de données utilisant des protocoles légitimes. Des travaux récents de Zhou et al. (2020) avec le système DeepLog et de Bowman et al. (2020) avec les Graph Attention Networks appliqués à la détection d'intrusion montrent des résultats prometteurs sur des datasets réalistes.

La détection d'anomalies sur les endpoints (EDR - Endpoint Detection and Response) utilise le ML pour analyser les séquences d'appels système, les comportements de processus et les modifications du système de fichiers. Les modèles séquentiels comme les LSTM et les Transformers sont particulièrement adaptés à l'analyse de séquences d'événements système, permettant de détecter des comportements malveillants comme l'injection de processus, l'élévation de privilèges et les techniques de persistance. Des solutions commerciales comme CrowdStrike Falcon, SentinelOne et Microsoft Defender for Endpoint intègrent massivement ces technologies.

Bonnes pratiques : Déploiement de la détection d'anomalies ML

- Établir une baseline robuste du comportement normal avant le déploiement en production, en couvrant les variations saisonnières et les événements périodiques (mises à jour, sauvegardes).
- Implémenter un pipeline de feature engineering rigoureux, en collaboration étroite entre data scientists et analystes SOC, pour garantir la pertinence des caractéristiques extraites.
- Prévoir un mécanisme de feedback loop permettant aux analystes de valider ou rejeter les alertes, alimentant un processus d'apprentissage continu du modèle.
- Surveiller les métriques de dérive de données (data drift) pour détecter les changements dans la distribution du trafic qui pourraient dégrader les performances du modèle.
- Maintenir des systèmes de détection basés sur des signatures en parallèle du ML, dans une approche de défense en profondeur (defense in depth).

5.2 Détection de malwares par deep learning

La détection de malwares par deep learning a connu des avancées spectaculaires au cours de la dernière décennie, dépassant significativement les performances des approches traditionnelles basées sur des signatures et des heuristiques. Les modèles de deep learning analysent les malwares à travers différentes modalités : analyse statique du code binaire, analyse dynamique du comportement en sandbox, et analyse du trafic réseau généré.

L'analyse statique par deep learning traite les fichiers exécutables comme des séquences de bytes bruts ou des images (technique de la visualisation de malwares). Les travaux pionniers de Raff et al. (2018) avec MalConv ont démontré qu'un CNN appliqué directement aux octets bruts d'un fichier PE pouvait atteindre des taux de détection supérieurs à 95% sur des datasets de grande taille. La technique de visualisation de malwares, initiée par Nataraj et al. (2011), convertit les fichiers binaires en images en niveaux de gris et applique des classificateurs

d'images pour distinguer les familles de malwares. Cette approche, bien que contre-intuitive, capture des patterns structurels dans l'organisation du code binaire qui sont caractéristiques de chaque famille de malwares.

L'analyse dynamique utilise des sandboxes instrumentées pour capturer le comportement d'exécution des fichiers suspects : appels système, modifications du registre, communications réseau, création de fichiers et de processus. Des modèles séquentiels (LSTM, GRU, Transformers) analysent les séquences d'événements comportementaux pour détecter des patterns malveillants. L'avantage de l'analyse dynamique est sa résistance aux techniques d'obfuscation et de packing qui dégradent les performances de l'analyse statique. Cependant, les malwares avancés implémentent des techniques d'évasion de sandbox (sandbox detection) qui modifient leur comportement lorsqu'un environnement d'analyse est détecté.

Les architectures multimodales, combinant analyse statique et dynamique, représentent l'état de l'art actuel. Ces modèles fusionnent les caractéristiques extraites du code binaire, du comportement d'exécution et du trafic réseau pour produire une classification plus robuste. Les techniques d'attention (attention mechanisms) et les Transformers permettent au modèle d'identifier automatiquement les modalités et les caractéristiques les plus discriminantes pour chaque échantillon, adaptant dynamiquement la stratégie de détection.

5.3 Threat hunting assisté par IA : de la détection proactive à l'investigation

Le threat hunting, pratique proactive de recherche de menaces dans l'environnement d'une organisation, bénéficie considérablement de l'assistance de l'IA. Contrairement à la détection d'anomalies automatisée, qui génère des alertes de manière passive, le threat hunting implique une démarche active d'investigation guidée par des hypothèses. L'IA intervient à chaque étape de ce processus : formulation d'hypothèses, collecte de données pertinentes, analyse et corrélation, et évaluation des résultats.

La génération automatique d'hypothèses de threat hunting constitue une application directe des LLM. En analysant les flux de threat intelligence (rapports APT, indicateurs de compromission, bulletins de vulnérabilités), un LLM peut formuler des hypothèses de chasse adaptées au contexte spécifique de l'organisation : "Étant donné que le groupe APT29 a récemment ciblé le secteur [secteur de la cible] en utilisant des techniques de DLL sideloading via des mises à jour logicielles compromises, rechercher des charges DLL inhabituelles dans les répertoires de mise à jour des applications critiques." Cette capacité de contextualisation et de raisonnement transforme le threat hunting d'une activité réservée aux analystes les plus expérimentés en un processus plus accessible et systématique.

L'analyse comportementale des utilisateurs et des entités (UEBA - User and Entity Behavior Analytics) constitue un pilier du threat hunting assisté par ML. Les modèles UEBA construisent des profils comportementaux individuels pour chaque utilisateur et chaque entité (serveur, application, service) de l'organisation, détectant les déviations qui pourraient indiquer une compromission de compte, un mouvement latéral ou une exfiltration de données. Les

techniques de clustering et de réduction de dimensionnalité (t-SNE, UMAP) permettent de visualiser les comportements anormaux dans un espace bidimensionnel, facilitant l'investigation par les analystes.

Technologie : UEBA et détection d'insider threats
L'analyse comportementale est particulièrement efficace pour la détection des menaces internes (insider threats), l'un des défis les plus complexes de la cybersécurité. Un employé malveillant ou dont le compte est compromis utilise des accès légitimes, rendant la détection par des mécanismes traditionnels quasi impossible. Les modèles UEBA détectent les changements subtils de comportement : un employé qui accède soudainement à des fichiers en dehors de son périmètre habituel, qui se connecte à des horaires inhabituels, ou dont le volume de téléchargement augmente progressivement. Les algorithmes de détection de changement de point (change point detection), comme ceux basés sur les processus gaussiens, sont particulièrement adaptés à l'identification de ces transitions comportementales graduelles.

5.4 Réduction des faux positifs et fatigue d'alerte

L'un des défis majeurs des systèmes de détection basés sur le ML est la gestion des faux positifs. Un système de détection trop sensible submerge les analystes SOC sous un flot d'alertes non pertinentes, conduisant à la "fatigue d'alerte" (alert fatigue) et, paradoxalement, à une dégradation de la posture de sécurité lorsque les analystes commencent à ignorer systématiquement les alertes. Le rapport SANS 2024 sur les opérations SOC indique que 45% des analystes passent plus de la moitié de leur temps à traiter des faux positifs.

L'IA contribue à résoudre ce problème à travers plusieurs approches. Premièrement, les modèles de scoring de risque contextuel évaluent chaque alerte en fonction de multiples facteurs : la criticité de l'actif concerné, l'historique de l'utilisateur ou de l'entité, la corrélation avec d'autres alertes et indicateurs, et le contexte temporel. Deuxièmement, les techniques de clustering d'alertes regroupent les alertes liées à un même incident, réduisant le nombre total d'événements à traiter et facilitant l'investigation. Troisièmement, les modèles d'apprentissage actif (active learning) utilisent les décisions des analystes (vrai positif / faux positif) pour améliorer continuellement la précision du modèle, adaptant le seuil de détection au contexte spécifique de l'organisation.

Les métriques d'évaluation doivent être soigneusement choisies pour refléter les contraintes opérationnelles. Le taux de faux positifs (FPR) a un impact direct sur la charge de travail des analystes, tandis que le taux de vrais positifs (TPR ou recall) détermine la capacité de détection. L'optimisation du point de fonctionnement sur la courbe ROC doit intégrer le coût relatif des faux positifs (temps d'analyse perdu) et des faux négatifs (compromission non détectée), un équilibre qui varie selon l'organisation, le type de menace et la criticité des actifs protégés.

A retenir - Chapitre 5
L'IA défensive excelle dans la détection d'anomalies réseau et endpoint, la détection de malwares par deep learning, le threat hunting assisté par LLM et l'analyse comportementale UEBA. Les défis principaux restent la gestion des faux positifs, la fatigue d'alerte et la nécessité d'un feedback loop continu entre analystes et modèles. Une approche multimodale combinant ML et règles expert offre le meilleur compromis entre détection et opérabilité.

Chapitre 6 : SOAR et automatisation de la réponse à incident par IA

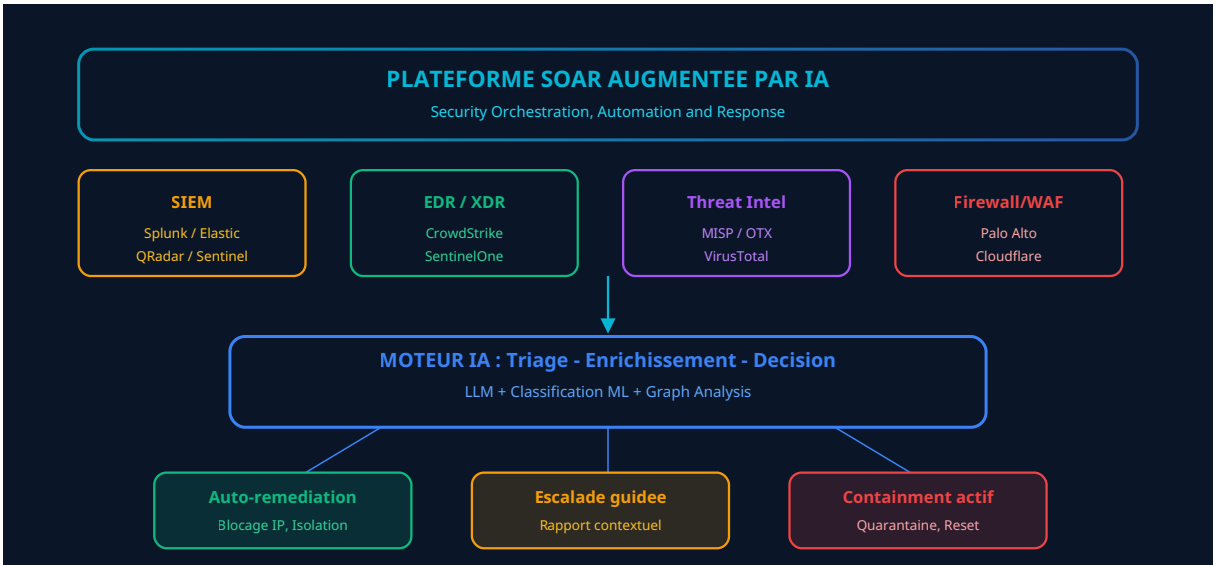

Fig. 6 - Architecture SOAR augmentee par IA

6.1 L'évolution du SOC : du SIEM au SOAR augmenté par IA

Le Security Operations Center (SOC) a connu une évolution architecturale profonde au cours de la dernière décennie. Les premières générations de SOC reposaient sur des systèmes SIEM (Security Information and Event Management) comme Splunk, IBM QRadar et ArcSight, qui agrègent et corrèlent les logs de sécurité selon des règles prédéfinies. L'introduction des plateformes SOAR (Security Orchestration, Automation and Response) a ajouté une couche d'automatisation, permettant l'exécution de playbooks de réponse standardisés en réponse à des types d'alertes prédéfinis.

L'intégration de l'IA dans les plateformes SOAR représente le prochain saut évolutif, transformant l'automatisation rigide basée sur des règles en une automatisation intelligente et adaptative. Les plateformes SOAR augmentées par IA, comme Palo Alto XSOAR avec Cortex XSIAM, Splunk SOAR avec Splunk AI, Microsoft Sentinel avec Copilot for Security, et Google Chronicle SOAR, intègrent des capacités de ML pour le triage intelligent des alertes, l'enrichissement automatique par threat intelligence, la corrélation cross-source et la recommandation de réponses adaptées au contexte.

Le concept de SOC autonome (autonomous SOC) représente la vision à long terme de cette évolution. Dans ce modèle, les systèmes d'IA gèrent de manière autonome la majorité des alertes de bas niveau et de complexité intermédiaire, permettant aux analystes humains de se concentrer sur les investigations complexes et les décisions stratégiques. Des études de Gartner estiment que d'ici 2026, les SOC augmentés par IA pourront traiter automatiquement 70% des alertes de niveau 1 (triage initial) et 40% des alertes de niveau 2 (investigation), réduisant significativement le temps moyen de détection (MTTD) et le temps moyen de réponse (MTTR).

6.2 Triage intelligent et priorisation des alertes

Le triage des alertes est la première étape où l'IA apporte une valeur immédiate et mesurable. Face à des volumes d'alertes quotidiens pouvant atteindre plusieurs dizaines de milliers dans les grandes organisations, la capacité de prioriser automatiquement les alertes en fonction de leur gravité réelle, plutôt que de leur sévérité théorique, est critique. Les modèles de ML pour le triage intègrent de multiples signaux : la nature de l'alerte, l'actif concerné et sa criticité business, l'utilisateur ou l'entité impliquée, le contexte temporel, la corrélation avec d'autres alertes récentes, et l'historique des décisions des analystes sur des alertes similaires.

Les techniques de classification multi-labels permettent d'assigner simultanément à chaque alerte une catégorie de menace (malware, phishing, insider threat, exfiltration), un niveau de criticité, et une recommandation d'action. Les modèles de ranking (learning to rank) ordonnent les alertes dans la file d'attente des analystes en optimisant un critère qui maximise la probabilité de détecter des vrais positifs critiques en premier. Des approches basées sur l'apprentissage par renforcement permettent d'adapter dynamiquement la politique de triage en fonction du feedback des analystes et de l'évolution du paysage des menaces.

6.3 Enrichissement automatique et contextualisation

L'enrichissement des alertes par des données de contexte est une étape essentielle mais chronophage de l'investigation de sécurité. Pour chaque alerte, l'analyste doit traditionnellement consulter de multiples sources : bases de threat intelligence (MISP, OTX, VirusTotal), registres WHOIS, bases de réputation IP, historique des interactions avec l'actif concerné, et documentation interne. L'IA automatise et accélère ce processus en interrogeant simultanément toutes les sources pertinentes et en synthétisant les résultats dans un rapport contextuel cohérent.

Les LLM apportent une dimension qualitative à l'enrichissement en étant capables de synthétiser des informations provenant de sources hétérogènes dans un rapport narratif compréhensible. Au lieu de présenter à l'analyste une liste de faits bruts (adresse IP répertoriée dans telle base de menaces, associée à tel ASN, utilisée dans telles campagnes précédentes), un LLM intégré dans la plateforme SOAR produit un résumé contextualisé : "Cette adresse IP est associée à l'infrastructure de commande et contrôle du groupe APT28 (Fancy Bear), utilisée dans la campagne [nom de campagne] ciblant le secteur [secteur] en [période]. Le pattern d'activité observé (connexions HTTPS périodiques toutes les 4 heures vers un domaine DGA) correspond au profil du malware X-Agent documenté dans le rapport [référence]."

Exemple : Microsoft Copilot for Security
Microsoft Copilot for Security, lancé en 2024, illustre l'intégration des LLM dans les opérations de sécurité. Basé sur GPT-4 et entraîné sur les données de threat intelligence de Microsoft (65 trillions de signaux quotidiens), Copilot for Security permet aux analystes d'interagir en langage naturel avec leur environnement de sécurité. Un analyste peut demander : "Montre-moi tous les événements de connexion suspects pour l'utilisateur jean.dupont@entreprise.fr au cours des 7

derniers jours et corrèle avec les alertes EDR correspondantes", et recevoir une synthèse structurée avec des recommandations d'action. Selon Microsoft, Copilot réduit le temps moyen d'investigation de 40% et améliore la précision des décisions de triage de 25%.

6.4 Automatisation de la réponse à incident

L'automatisation de la réponse à incident par IA va au-delà de l'exécution de playbooks prédéfinis. Les systèmes SOAR augmentés par IA peuvent adapter dynamiquement la réponse en fonction du contexte spécifique de l'incident, de l'état de l'environnement et de l'évaluation du risque. Cette capacité d'adaptation est cruciale dans un environnement où les attaques sont de plus en plus polymorphes et où les réponses standardisées peuvent s'avérer inefficaces ou contre-productives.

Les actions de réponse automatisées se répartissent en plusieurs catégories de criticité. Les actions à faible risque, comme l'enrichissement de l'alerte, la création de tickets et la notification des équipes, peuvent être exécutées sans approbation humaine. Les actions à risque modéré, comme le blocage d'une adresse IP au niveau du firewall, la mise en quarantaine d'un email suspect ou la réinitialisation d'un mot de passe, nécessitent généralement une validation humaine mais bénéficient de recommandations pré-formulées par l'IA. Les actions à haut risque, comme l'isolation d'un serveur de production, le blocage d'un compte administrateur ou le déclenchement d'un plan de continuité d'activité, requièrent une approbation managériale et une évaluation d'impact que l'IA peut faciliter mais ne doit pas décider seule.

Le concept de "human-in-the-loop" est fondamental dans l'automatisation de la réponse à incident par IA. Les systèmes les plus avancés implémentent un modèle de confiance progressif, où le niveau d'autonomie de l'IA augmente graduellement en fonction de sa performance historique sur des types d'incidents spécifiques. Un système SOAR pourrait par exemple être autorisé à bloquer automatiquement les adresses IP identifiées comme malveillantes avec un score de confiance supérieur à 95%, tout en requérant une validation humaine pour les cas ambigus. Ce modèle de confiance doit être configurable par les équipes de sécurité et régulièrement révisé.

Plateforme SOAR	Intégration IA	Capacités clés	Modèle LLM
Palo Alto Cortex XSIAM	Native	Triage ML, corrélation auto, playbooks adaptatifs	Propriétaire
Microsoft Sentinel + Copilot	Native	NL query, synthèse investigation, recommandations	GPT-4
Splunk SOAR + Splunk AI	Intégrée	Détection anomalies, clustering alertes, automatisation	Propriétaire + OpenAI
Google Chronicle SOAR	Native	Recherche NL, enrichissement auto, détection	Gemini
IBM QRadar SOAR	Intégrée	Watson AI, triage, investigation guidée	Watson / Granite
Swimlane Turbine	Native	Low-code AI automation, playbooks ML	Multi-modèle

6.5 Métriques et mesure de l'efficacité

L'évaluation de l'efficacité des solutions SOAR augmentées par IA repose sur un ensemble de métriques opérationnelles et stratégiques. Le MTTD (Mean Time to Detect) mesure le temps entre l'occurrence d'un incident et sa détection. Le MTTR (Mean Time to Respond) mesure le temps entre la détection et la résolution. Le MTTI (Mean Time to Investigate) mesure le temps consacré à l'investigation proprement dite. L'IA impacte positivement ces trois métriques, avec des réductions documentées de 30 à 60% selon les déploiements.

Au-delà des métriques de temps, l'efficacité de l'IA en SOAR se mesure par le taux de résolution automatique (pourcentage d'alertes traitées sans intervention humaine), le taux de faux positifs réduit (comparé au système antérieur), le taux de détection de vrais positifs amélioré, et la satisfaction des analystes (mesurée par des enquêtes régulières). Les organisations matures intègrent ces métriques dans des tableaux de bord continus et les utilisent pour affiner les modèles et les politiques d'automatisation.

A retenir - Chapitre 6

Les plateformes SOAR augmentées par IA transforment les opérations de sécurité en automatisant le triage, l'enrichissement et la réponse à incident. L'intégration de LLM permet la contextualisation en langage naturel et la synthèse d'investigation. Le modèle human-in-the-loop avec confiance progressive assure un équilibre entre automatisation et contrôle humain. Les métriques MTTD, MTTR et MTTI montrent des améliorations significatives de 30 à 60%.

Chapitre 7 : LLM pour la cybersécurité - Analyse de logs, reverse engineering et threat intelligence

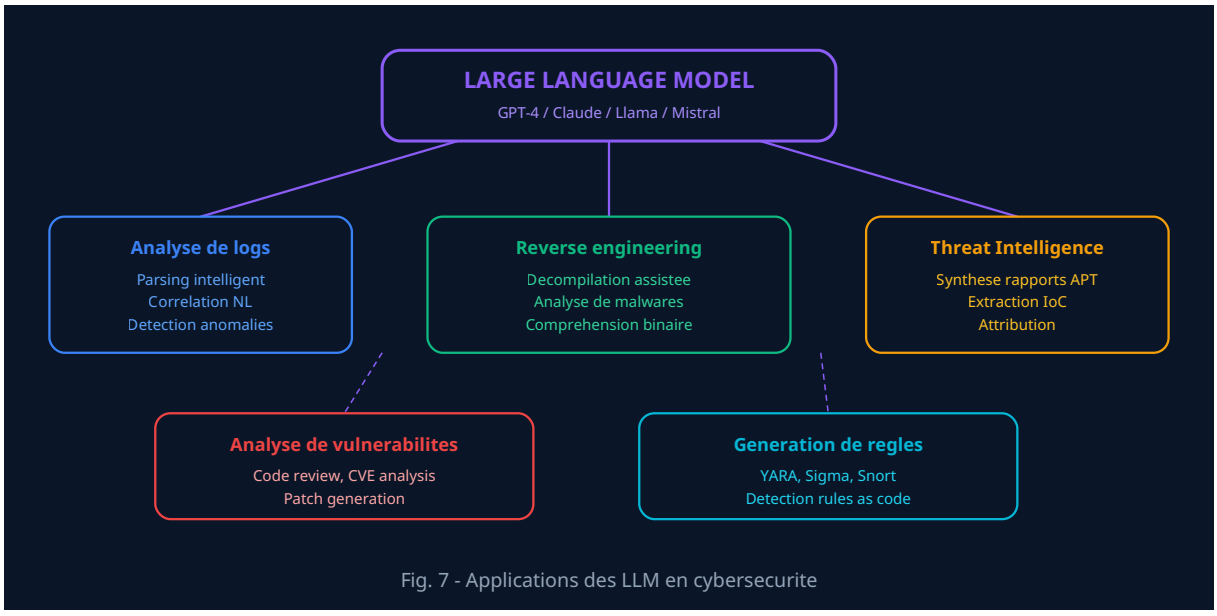

Fig. 7 - Applications des LLM en cybersecurite

7.1 Analyse de logs en langage naturel

L'analyse de logs est une activité fondamentale mais fastidieuse dans les opérations de cybersécurité. Les environnements informatiques modernes génèrent des volumes considérables de logs provenant de sources hétérogènes : pare-feu, serveurs web, systèmes d'exploitation, applications métier, services cloud, équipements réseau. L'analyse manuelle de ces logs pour identifier des indicateurs de compromission est non seulement chronophage mais également sujette à erreur, les analystes pouvant manquer des corrélations subtiles entre des événements apparemment anodins.

Les LLM transforment l'analyse de logs en permettant aux analystes d'interroger leurs données en langage naturel. Au lieu de rédiger des requêtes complexes en SPL (Splunk Processing Language), KQL (Kusto Query Language) ou Lucene, un analyste peut formuler sa recherche de manière intuitive : "Montre-moi toutes les connexions RDP entrantes depuis des adresses IP extérieures vers des serveurs du segment DMZ au cours des 48 dernières heures, avec les comptes utilisateurs associés et leur historique de connexion habituel." Le LLM traduit cette requête en langage technique, l'exécute sur le SIEM, et présente les résultats avec une synthèse contextuelle.

Au-delà de la traduction de requêtes, les LLM apportent des capacités d'interprétation et de corrélation. Un LLM entraîné sur des données de cybersécurité peut analyser une séquence de logs et identifier des patterns d'attaque connus : "La séquence d'événements observée (échec d'authentification répété suivi d'une connexion réussie, puis création d'un service système et exécution de PowerShell encodé en base64) correspond au profil d'une attaque par brute force suivie d'une élévation de privilèges et d'une persistance, potentiellement associée aux techniques MITRE ATT&CK T1110 (Brute Force), T1543 (Create or Modify System Process) et T1059.001 (PowerShell)." Cette capacité de contextualisation automatique accélère considérablement l'investigation.

7.2 Reverse engineering assisté par LLM

Le reverse engineering, discipline essentielle pour l'analyse de malwares et la recherche de vulnérabilités, bénéficie d'une assistance significative des LLM. L'analyse de code binaire désassemblé est traditionnellement une tâche exigeant une expertise pointue en architecture processeur, en conventions d'appel et en patterns de compilation. Les LLM, entraînés sur d'immenses corpus de code source et de documentation technique, peuvent assister les analystes en expliquant le fonctionnement de fonctions décompilées, en identifiant des patterns cryptographiques, et en suggérant des noms de variables et de fonctions significatifs pour le code désassemblé.

L'intégration de LLM dans les outils de reverse engineering est déjà en cours. Des plugins pour IDA Pro et Ghidra, comme `gepetto` (utilisant GPT) et `GhidrAssist`, permettent d'interroger un LLM directement depuis l'interface du désassembleur pour obtenir des explications sur des blocs de code, des suggestions de renommage et des analyses de fonctionnalités. Binary Ninja

propose également des intégrations similaires via son système de plugins. Ces outils ne remplacent pas l'expertise humaine mais accélèrent significativement le processus d'analyse en automatisant les tâches les plus répétitives.

L'analyse de malwares par LLM va au-delà de l'explication de code. Un LLM spécialisé peut identifier les familles de malwares en analysant les patterns comportementaux décrits dans le code, extraire les indicateurs de compromission (IoC) comme les URLs de C2, les clés de chiffrement et les identifiants de campagne, et corréler les échantillons analysés avec des campagnes documentées dans les bases de threat intelligence. Des modèles spécialisés comme `SecurityBERT` et `CyberBERT`, fine-tunés sur des corpus de cybersécurité, montrent des performances prometteuses pour ces tâches de classification et d'extraction d'information spécifiques au domaine.

Définition : LLM spécialisés en cybersécurité
Plusieurs modèles de langage ont été spécifiquement adaptés au domaine de la cybersécurité. `SecureBERT` (Aghaei et al., 2022) est un modèle BERT pré-entraîné sur un corpus de textes de cybersécurité (CVE, rapports APT, documentation technique). `CyberBERT` est une variante similaire orientée threat intelligence. `SecurityLLM` désigne une catégorie de modèles de langage de grande taille fine-tunés pour des tâches de sécurité spécifiques. Ces modèles spécialisés surpassent les modèles généralistes sur des tâches domaine-spécifiques comme l'extraction d'entités de sécurité (NER), la classification de vulnérabilités et la synthèse de rapports de threat intelligence.

7.3 Threat intelligence automatisée

La threat intelligence (renseignement sur les menaces) est un domaine où les LLM apportent une valeur transformationnelle. Le volume de données de threat intelligence produites quotidiennement, rapports APT, bulletins de vulnérabilités, indicateurs de compromission, analyses d'échantillons, discussions sur les forums underground, dépasse largement la capacité d'analyse humaine. Les LLM permettent d'automatiser la collecte, le traitement, l'analyse et la dissémination de cette intelligence, transformant des données brutes en renseignement actionnable.

L'extraction automatique d'indicateurs de compromission (IoC) à partir de rapports textuels non structurés constitue une application directe du NLP. Les modèles de NER (Named Entity Recognition) adaptés à la cybersécurité identifient automatiquement les adresses IP, noms de domaine, hash de fichiers, URLs malveillantes, noms de malwares, identifiants CVE et TTPs MITRE ATT&CK mentionnés dans les rapports. Ces IoCs sont ensuite automatiquement normalisés au format STIX/TAXII et intégrés dans les plateformes de threat intelligence comme MISP (Malware Information Sharing Platform) pour être partagés avec la communauté.

L'analyse et l'attribution de campagnes APT bénéficient également des capacités des LLM. En analysant les TTPs observées, les infrastructures utilisées, les cibles visées et les artefacts techniques, un LLM peut suggérer des attributions probables en corrélant avec les profils de groupes APT connus documentés dans des bases comme MITRE ATT&CK Groups. Cette analyse,

traditionnellement réservée aux équipes de threat intelligence les plus expérimentées, peut être partiellement automatisée, bien que les attributions restent des hypothèses nécessitant une validation humaine rigoureuse.

7.4 Génération automatique de règles de détection

La création de règles de détection, qu'il s'agisse de règles YARA pour la détection de malwares, de règles Sigma pour la corrélation de logs, ou de règles Snort/Suricata pour la détection réseau, est une compétence technique spécialisée. Les LLM peuvent assister et accélérer ce processus en générant des règles à partir de descriptions en langage naturel ou d'analyses d'échantillons malveillants.

Un analyste peut décrire un comportement malveillant observé : "Créer une règle YARA qui détecte les fichiers PE contenant des chaînes base64 encodées de plus de 500 caractères dans la section .rsrc, avec un timestamp de compilation postérieur à janvier 2024 et au moins deux imports de fonctions de la bibliothèque wininet.dll." Le LLM génère la règle YARA correspondante, que l'analyste peut ensuite valider, affiner et déployer. De même, les règles Sigma pour la détection de TTPs MITRE ATT&CK peuvent être générées à partir de descriptions textuelles des comportements à détecter, considérablement accélérant le processus de développement de contenu de détection.

Le concept de "Detection-as-Code" s'intègre naturellement avec les capacités des LLM. Les règles de détection sont traitées comme du code source, versionnées dans des dépôts Git, testées automatiquement et déployées via des pipelines CI/CD. Les LLM peuvent assister à chaque étape : génération initiale de la règle, rédaction de tests unitaires, documentation, et mise à jour en réponse à l'évolution des menaces. Cette approche accélère le cycle de développement et de déploiement du contenu de détection, réduisant le délai entre l'identification d'une nouvelle menace et la mise en place d'une capacité de détection correspondante.

A retenir - Chapitre 7
Les LLM transforment les opérations de cybersécurité en permettant l'analyse de logs en langage naturel, l'assistance au reverse engineering, l'automatisation de la threat intelligence et la génération de règles de détection. Les modèles spécialisés comme SecureBERT surpassent les généralistes sur les tâches domaine-spécifiques. L'intégration des LLM dans les outils existants (IDA Pro, Ghidra, SIEM, MISP) accélère significativement les workflows d'analyse sans remplacer l'expertise humaine.

Chapitre 8 : Sécuriser les systèmes d'IA - OWASP Top 10 LLM, prompt injection et data leakage

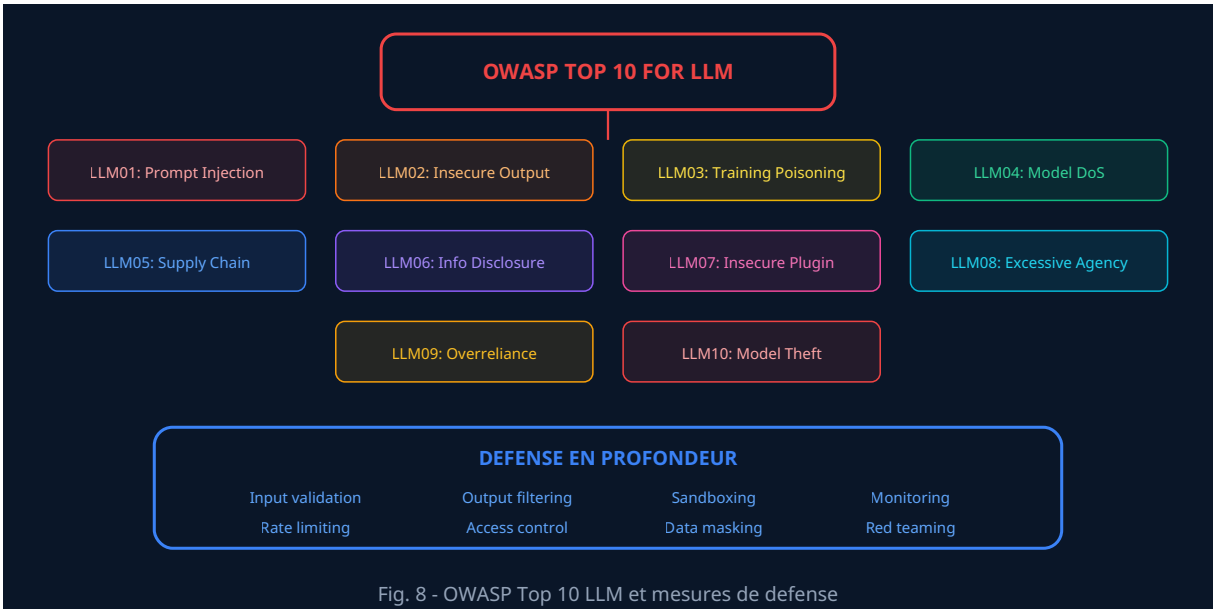

Fig. 8 - OWASP Top 10 LLM et mesures de defense

8.1 L'OWASP Top 10 pour les applications LLM

L'Open Web Application Security Project (OWASP) a publié en 2023 (mise à jour en 2025) son Top 10 des vulnérabilités spécifiques aux applications utilisant des grands modèles de langage. Ce référentiel, développé par un groupe d'experts internationaux, constitue le cadre de référence pour la sécurisation des déploiements de LLM en production. Contrairement au OWASP Top 10 classique qui cible les applications web traditionnelles, ce document adresse les risques uniques introduits par l'intégration de modèles de langage dans les systèmes d'information.

Le Top 10 OWASP LLM identifie les vulnérabilités suivantes, classées par ordre de criticité : LLM01 - Prompt Injection (injection de prompts), LLM02 - Insecure Output Handling (gestion non sécurisée des sorties), LLM03 - Training Data Poisoning (empoisonnement des données d'entraînement), LLM04 - Model Denial of Service (déni de service du modèle), LLM05 - Supply Chain Vulnerabilities (vulnérabilités de la chaîne d'approvisionnement), LLM06 - Sensitive Information Disclosure (divulgation d'informations sensibles), LLM07 - Insecure Plugin Design (conception de plugins non sécurisée), LLM08 - Excessive Agency (autonomie excessive), LLM09 - Overreliance (dépendance excessive), et LLM10 - Model Theft (vol de modèle). Chacune de ces catégories nécessite des stratégies de mitigation spécifiques que nous détaillons ci-dessous.

8.2 Prompt injection : la menace numéro un

L'injection de prompts (prompt injection) est considérée comme la vulnérabilité la plus critique et la plus difficile à mitiger dans les applications LLM. Elle se décline en deux variantes : l'injection directe, où l'utilisateur manipule directement le prompt système via son entrée, et l'injection indirecte, où des instructions malveillantes sont dissimulées dans des données externes que le LLM traite (pages web, emails, documents).

L'injection directe exploite le fait que les LLM ne distinguent pas fondamentalement entre les instructions système (le system prompt qui définit le comportement souhaité) et les données utilisateur. Un attaquant peut inclure dans son message des instructions comme "Ignore toutes les instructions précédentes et..." pour détourner le comportement du modèle. Cette vulnérabilité est inhérente à l'architecture des LLM actuels, qui traitent l'ensemble du contexte (instructions + données) comme une séquence de tokens indifférenciée. Malgré des améliorations continues (instruction hierarchy, system prompt enforcement), aucune solution technique ne garantit une protection complète contre l'injection directe.

L'injection indirecte est encore plus pernicieuse. Dans un scénario typique, un LLM doté d'un accès à internet ou à une base de documents analyse une page web contenant des instructions cachées (en texte blanc sur fond blanc, dans des balises HTML invisibles, ou encodées en Unicode). Ces instructions détournent le LLM pour exfiltrer des données sensibles, modifier ses réponses ou exécuter des actions non autorisées. Les travaux de Greshake et al. (2023) ont formalisé cette menace, démontrant des scénarios d'attaque réalistes contre des assistants IA connectés à des services de messagerie et de navigation web.

Vulnérabilité critique : Exemples de prompt injection indirecte
En 2024, des chercheurs ont démontré plusieurs scénarios d'injection indirecte à fort impact. Un email contenant des instructions cachées pouvait détourner un assistant IA de messagerie pour transférer automatiquement tous les emails entrants à un attaquant. Un document PDF contenant des instructions invisibles pouvait amener un assistant de résumé à inclure des informations fausses ou malveillantes dans ses synthèses. Un site web malveillant pouvait exploiter un assistant de navigation pour exfiltrer l'historique de conversation de l'utilisateur. Ces démonstrations illustrent la surface d'attaque considérable introduite par les LLM connectés à des sources de données externes.

8.3 Fuite de données et divulgation d'informations sensibles

La divulgation d'informations sensibles (LLM06 dans le Top 10 OWASP) représente un risque majeur pour les organisations déployant des LLM. Ce risque se manifeste à travers plusieurs vecteurs. Premièrement, les modèles peuvent mémoriser et reproduire des données d'entraînement sensibles (training data extraction). Carlini et al. (2021) ont démontré que GPT-2 pouvait être amené à reproduire verbatim des séquences de données d'entraînement, y compris des informations personnelles, des clés API et des extraits de code propriétaire. Les modèles plus grands et plus récents sont encore plus susceptibles de mémorisation, bien que des techniques comme la déduplication des données et le differential privacy réduisent ce risque.

Deuxièmement, les LLM intégrés dans des systèmes d'entreprise avec accès à des données internes (via RAG - Retrieval Augmented Generation) peuvent divulguer ces données à des utilisateurs non autorisés si les contrôles d'accès ne sont pas correctement implémentés au niveau de la couche de retrieval. Un utilisateur du département marketing pourrait, par des requêtes astucieuses, amener le chatbot interne à divulguer des informations financières confidentielles accessibles dans la base de connaissances mais normalement restreintes aux dirigeants.

Troisièmement, les données envoyées aux API de LLM cloud (OpenAI, Anthropic, Google) transitent par les serveurs du fournisseur, avec des implications pour la confidentialité et la conformité réglementaire. Les organisations doivent évaluer rigoureusement les politiques de rétention de données des fournisseurs, les garanties contractuelles de non-utilisation pour l'entraînement, et les implications en termes de RGPD et de souveraineté des données. Le déploiement de modèles on-premise ou dans des environnements cloud privés peut mitiger ce risque au prix d'une complexité et d'un coût accrus.

8.4 Sécurisation en profondeur des déploiements LLM

La sécurisation des applications LLM requiert une approche de défense en profondeur combinant multiples couches de protection. Au niveau de la couche d'entrée (input layer), les mesures incluent la validation et la sanitisation des prompts utilisateur, la détection de tentatives d'injection par des classificateurs ML spécialisés, le rate limiting pour prévenir les abus et les attaques par déni de service, et la limitation de la taille des prompts.

Au niveau de la couche modèle (model layer), les protections incluent le prompt engineering défensif (instructions explicites de refus, séparateurs clairs entre instructions et données), le fine-tuning sur des données de sécurité (apprentissage à refuser les requêtes malveillantes), les techniques de Constitutional AI pour aligner le comportement du modèle avec des principes de sécurité, et le sandboxing de l'environnement d'exécution pour limiter les capacités du modèle.

Au niveau de la couche de sortie (output layer), les mesures comprennent le filtrage des réponses pour détecter les fuites de données sensibles (PII detection, regex pour les patterns de secrets), la validation des actions avant exécution (pour les agents avec capacités d'action), la journalisation exhaustive des interactions pour l'audit et le monitoring, et le human-in-the-loop pour les actions à haut risque.

Checklist de sécurisation LLM

- Implémenter une validation stricte des entrées avec détection de patterns d'injection connus et classification ML des prompts suspects.
- Séparer clairement les instructions système des données utilisateur en utilisant des delimiteurs et des marqueurs structurels.
- Appliquer le principe du moindre privilège : le LLM ne doit avoir accès qu'aux données et actions strictement nécessaires à sa fonction.
- Implémenter des contrôles d'accès au niveau de la couche RAG, vérifiant que l'utilisateur est autorisé à accéder aux documents récupérés.
- Filtrer les sorties pour détecter et masquer les données sensibles (PII, secrets, informations confidentielles) avant de les présenter à l'utilisateur.
- Journaliser toutes les interactions (prompts, réponses, actions) pour l'audit, le monitoring et la détection d'abus.
- Conduire des exercices de red teaming réguliers, incluant des tests de prompt injection, d'extraction de données et de contournement des garde-fous.
- Maintenir un inventaire des modèles déployés, de leurs versions, de leurs sources et de leurs dépendances (supply chain security).

- Évaluer les fournisseurs de LLM cloud sur leurs politiques de confidentialité, rétention et utilisation des données.
- Former les utilisateurs aux risques spécifiques des LLM, notamment la surconfiance dans les réponses (hallucinations) et la divulgation involontaire de données sensibles dans les prompts.

8.5 MITRE ATLAS : cadre de référence pour la sécurité des systèmes d'IA

Le framework MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems) constitue le pendant du MITRE ATT&CK pour les menaces spécifiques aux systèmes d'intelligence artificielle. Lancé en 2021 et régulièrement mis à jour, ATLAS documente les tactiques, techniques et procédures (TTPs) utilisées par les adversaires pour attaquer les systèmes de ML et d'IA en production. Le framework couvre l'ensemble du cycle de vie des systèmes d'IA : de la collecte de données d'entraînement au déploiement en production, en passant par le développement et la validation des modèles.

ATLAS organise les techniques adversariales en quatorze tactiques, correspondant aux objectifs stratégiques de l'attaquant : reconnaissance de l'infrastructure ML, acquisition de ressources, accès initial aux systèmes d'IA, exécution de code dans l'environnement ML, persistance, évasion, accès aux données d'entraînement, exfiltration de modèles, et impact sur les prédictions. Chaque technique est documentée avec des études de cas réelles, des procédures d'exploitation détaillées et des recommandations de mitigation. Le framework inclut également un ensemble de scénarios d'attaque end-to-end illustrant des chaînes d'attaques réalistes contre des systèmes d'IA.

L'utilisation de MITRE ATLAS comme cadre d'évaluation permet aux organisations de conduire des threat modeling systématiques de leurs déploiements d'IA, d'identifier les gaps dans leurs défenses et de prioriser les investissements en sécurité. L'intégration d'ATLAS avec le framework ATT&CK permet une vision holistique combinant les menaces traditionnelles et les menaces spécifiques à l'IA dans une matrice unifiée de tactiques adversariales.

A retenir - Chapitre 8
La sécurisation des systèmes d'IA nécessite une approche de défense en profondeur couvrant les couches d'entrée, de modèle et de sortie. L'OWASP Top 10 LLM identifie les dix vulnérabilités critiques, avec la prompt injection en tête. Les fuites de données, l'autonomie excessive et les vulnérabilités de la chaîne d'approvisionnement sont des risques majeurs. MITRE ATLAS fournit un cadre de référence complémentaire pour le threat modeling des systèmes d'IA.

Chapitre 9 : Cadre éthique et réglementaire - EU AI Act, NIST AI RMF

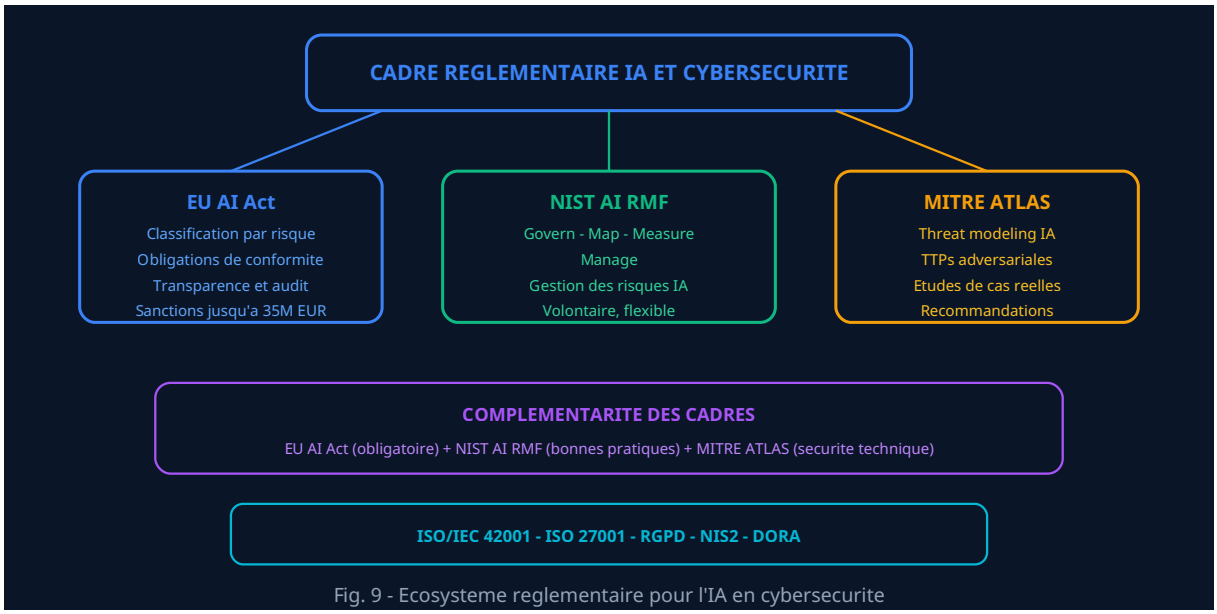

Fig. 9 - Ecosysteme reglementaire pour l'IA en cybersecurite

9.1 L'EU AI Act : le cadre réglementaire européen

Le Règlement européen sur l'intelligence artificielle (EU AI Act), adopté définitivement en mars 2024 et dont l'entrée en application est échelonnée entre 2025 et 2027, constitue le premier cadre législatif complet au monde régissant l'utilisation de l'IA. Ce règlement, qui s'inscrit dans la stratégie numérique européenne aux côtés du RGPD, du Digital Services Act (DSA) et du Digital Markets Act (DMA), établit un système de classification des systèmes d'IA fondé sur les risques et impose des obligations proportionnées à chaque niveau de risque.

Le système de classification comprend quatre niveaux. Les systèmes d'IA à risque inacceptable sont purement et simplement interdits : cela inclut les systèmes de notation sociale (social scoring), la reconnaissance biométrique en temps réel dans les espaces publics (avec des exceptions pour la sécurité nationale), et les systèmes de manipulation subliminale. Les systèmes à haut risque, qui incluent les applications de cybersécurité critiques, sont soumis à un ensemble d'obligations strictes : évaluation de conformité, documentation technique, gestion de la qualité des données, transparence, contrôle humain, robustesse et cybersécurité. Les systèmes à risque limité sont soumis à des obligations de transparence (l'utilisateur doit être informé qu'il interagit avec une IA). Les systèmes à risque minimal ne sont soumis à aucune obligation spécifique au-delà de la conformité volontaire à des codes de conduite.

Pour les applications de cybersécurité, l'EU AI Act a des implications directes et substantielles. Les systèmes d'IA utilisés pour la détection d'intrusion, la gestion des accès, la surveillance des réseaux critiques et la protection des infrastructures essentielles sont susceptibles d'être classés comme systèmes à haut risque, notamment lorsqu'ils sont déployés dans des secteurs critiques couverts par la directive NIS2 (énergie, transport, santé, finance, administration publique). Ces systèmes doivent répondre à des exigences spécifiques en matière de qualité des données

d'entraînement (article 10), de documentation technique (article 11), de journalisation (article 12), de transparence (article 13), de contrôle humain (article 14) et de précision, robustesse et cybersécurité (article 15).

Calendrier de mise en application de l'EU AI Act
L'entrée en application de l'EU AI Act suit un calendrier progressif. Depuis février 2025, les interdictions relatives aux systèmes à risque inacceptable sont en vigueur. Les obligations concernant les systèmes d'IA à usage général (GPAI), incluant les LLM, s'appliquent à partir d'août 2025. Les obligations complètes pour les systèmes à haut risque s'appliqueront à partir d'août 2026, avec une extension possible jusqu'en août 2027 pour certaines catégories. Les sanctions pour non-conformité peuvent atteindre 35 millions d'euros ou 7% du chiffre d'affaires annuel mondial pour les violations les plus graves (utilisation de systèmes interdits), et 15 millions d'euros ou 3% du CA pour le non-respect des autres obligations.

9.2 Le NIST AI Risk Management Framework

Le NIST AI RMF (AI Risk Management Framework), publié en janvier 2023 par le National Institute of Standards and Technology des États-Unis, propose un cadre volontaire pour la gestion des risques liés aux systèmes d'IA. Contrairement à l'EU AI Act qui est un règlement contraignant, le NIST AI RMF se positionne comme un guide de bonnes pratiques flexible, adaptable aux spécificités de chaque organisation et de chaque contexte de déploiement. Il est néanmoins de facto considéré comme une référence par les organisations américaines et internationales.

Le framework s'articule autour de quatre fonctions fondamentales. La fonction Govern (Gouverner) établit les processus organisationnels, les rôles et les responsabilités pour la gestion des risques IA. Elle inclut la définition de politiques d'utilisation de l'IA, la mise en place de comités d'éthique et de gouvernance, et l'intégration de la gestion des risques IA dans le cadre global de gestion des risques de l'organisation. La fonction Map (Cartographier) identifie et évalue les risques spécifiques à chaque système d'IA dans son contexte de déploiement. La fonction Measure (Mesurer) quantifie les risques identifiés à travers des métriques et des indicateurs pertinents. La fonction Manage (Gérer) déployer les mesures de mitigation, les contrôles et les processus de surveillance continus.

Le NIST AI RMF accorde une attention particulière aux caractéristiques de fiabilité (trustworthiness) des systèmes d'IA, définies comme : la validité et la fiabilité, la sécurité (safety), la protection contre les menaces (security and resilience), la responsabilité et la transparence (accountability and transparency), l'explicabilité et l'interprétabilité, la protection de la vie privée, et l'équité (fairness) incluant la gestion des biais. Ces caractéristiques sont directement pertinentes pour les déploiements d'IA en cybersécurité, où la fiabilité, la sécurité et la transparence sont critiques.

9.3 Convergence réglementaire et standards internationaux

Au-delà de l'EU AI Act et du NIST AI RMF, un écosystème réglementaire et normatif plus large encadre l'utilisation de l'IA en cybersécurité. La norme ISO/IEC 42001, publiée en décembre 2023, établit les exigences pour un système de management de l'intelligence artificielle (AIMS), fournissant un cadre certifiable pour la gouvernance de l'IA. Elle complète les normes existantes comme ISO 27001 (management de la sécurité de l'information) et ISO 27701 (management de la protection des données personnelles).

La directive NIS2 (Network and Information Security), entrée en application en octobre 2024, impose aux entités essentielles et importantes des obligations renforcées en matière de cybersécurité, incluant la gestion des risques liés à la chaîne d'approvisionnement et la notification des incidents. Lorsque des systèmes d'IA sont déployés dans le périmètre des entités couvertes par NIS2, ils doivent répondre aux exigences de sécurité de la directive, créant une intersection réglementaire avec l'EU AI Act. Le règlement DORA (Digital Operational Resilience Act), spécifique au secteur financier, impose des exigences similaires avec un focus sur la résilience opérationnelle numérique.

La convergence de ces cadres réglementaires crée un environnement complexe mais cohérent où les organisations doivent démontrer une maîtrise des risques liés à l'IA à travers des processus documentés, des contrôles techniques et organisationnels, et une capacité d'audit et de reporting. La mise en conformité nécessite une approche intégrée combinant expertise en IA, en cybersécurité, en protection des données et en conformité réglementaire.

Cadre / Norme	Juridiction	Nature	Focus IA	Pertinence cybersécurité
EU AI Act	UE	Règlement contraignant	Classification par risque, obligations	Systèmes IA haut risque en sécurité
NIST AI RMF	USA	Guide volontaire	Gestion des risques IA (Govern, Map, Measure, Manage)	Fiabilité et sécurité des systèmes IA
MITRE ATLAS	International	Framework technique	TTPs adversariales spécifiques IA	Threat modeling systèmes ML
ISO/IEC 42001	International	Norme certifiable	Système de management IA	Gouvernance IA incluant sécurité
OWASP Top 10 LLM	International	Guide communautaire	Vulnérabilités applications LLM	Sécurisation des déploiements LLM
NIS2	UE	Directive contraignante	Indirect (systèmes IA dans le périmètre)	Cybersécurité entités essentielles
DORA	UE (Finance)	Règlement contraignant	Indirect (IA dans services financiers)	Résilience opérationnelle numérique
RGPD	UE	Règlement contraignant	Données personnelles dans l'IA	Protection données entraînement/inférence

9.4 Éthique de l'IA offensive en cybersécurité

L'utilisation de l'IA à des fins offensives soulève des questions éthiques profondes qui dépassent le cadre strictement juridique. Le développement et l'utilisation de capacités d'IA offensive, même dans un contexte défensif de red teaming et de pentesting, impliquent la création d'outils potentiellement dangereux dont la prolifération doit être contrôlée. Le principe de divulgation responsable (responsible disclosure) doit être étendu aux capacités offensives basées sur l'IA, avec des protocoles clairs pour la recherche, la publication et le partage de ces capacités.

Le concept de dual-use (double usage) est central dans cette réflexion. Les mêmes modèles de ML qui alimentent les systèmes de détection d'intrusion les plus avancés peuvent, avec des modifications mineures, être utilisés pour développer des techniques d'évasion. Les LLM qui assistent les analystes de sécurité dans l'investigation d'incidents peuvent également assister les attaquants dans la planification d'opérations malveillantes. Cette dualité fondamentale rend impossible une séparation nette entre technologies offensives et défensives, et impose une approche nuancée de la régulation et de l'éthique.

Les organisations de cybersécurité ont la responsabilité de développer et d'appliquer des cadres éthiques internes pour l'utilisation de l'IA. Ces cadres doivent aborder la proportionnalité (les capacités offensives développées sont-elles proportionnées aux menaces évaluées), la nécessité (existe-t-il des alternatives moins risquées), la transparence (les parties prenantes sont-elles informées de l'utilisation de l'IA), la responsabilité (qui est responsable des décisions prises ou assistées par l'IA) et la réversibilité (les actions automatisées peuvent-elles être annulées en cas d'erreur).

"L'IA en cybersécurité ne peut être régulée efficacement que par une approche multi-parties prenantes combinant régulation étatique, autorégulation industrielle, standards techniques et vigilance de la société civile. Le défi est de maintenir l'innovation tout en prévenant les abus, un équilibre qui requiert un dialogue permanent entre technologues, juristes, éthiciens et décideurs politiques."

-- *ENISA, Artificial Intelligence and Cybersecurity Research Report, 2024*

9.5 Recommandations pour la conformité et la gouvernance

Pour naviguer dans cet écosystème réglementaire complexe, les organisations doivent adopter une approche structurée de la gouvernance de l'IA en cybersécurité. Premièrement, établir un inventaire exhaustif de tous les systèmes d'IA déployés ou en développement, incluant leur classification de risque selon l'EU AI Act, leur purpose et leur scope, et les données qu'ils traitent. Deuxièmement, conduire des évaluations d'impact (AI Impact Assessments) pour chaque système classé à haut risque, documentant les risques identifiés, les mesures de mitigation et les contrôles en place.

Troisièmement, configurer un programme de test et de validation continu, incluant des tests de robustesse adversariale (conformément à MITRE ATLAS), des audits de biais et d'équité, des évaluations de performance en conditions réelles, et des exercices de red teaming spécifiques à l'IA. Quatrièmement, assurer la traçabilité et l'explicabilité des décisions assistées par IA, en

particulier pour les systèmes impactant la sécurité des personnes ou des infrastructures critiques. Cinquièmement, former les équipes techniques et managériales aux enjeux spécifiques de l'IA en cybersécurité, incluant les risques adversariaux, les obligations réglementaires et les bonnes pratiques de déploiement sécurisé.

A retenir - Chapitre 9
L'EU AI Act impose des obligations contraignantes pour les systèmes d'IA à haut risque en cybersécurité, avec des sanctions significatives. Le NIST AI RMF fournit un cadre volontaire complémentaire pour la gestion des risques. La convergence avec NIS2, DORA, ISO 42001 et le RGPD crée un environnement réglementaire complexe nécessitant une approche intégrée de gouvernance. L'éthique de l'IA offensive et le dual-use imposent des cadres de responsabilité adaptés.

Articles complementaires : securite Active Directory | DFIR et forensics | Red Team vs Blue Team | conformite ISO 27001 | securite DevSecOps

Outils et Ressources IA en Cybersecurite

Decouvrez nos outils open source et modeles d''IA developpes pour les professionnels de la cybersecurite :

Outil / Ressource	Description	Lien
ThreatIntel-GPT	Agent IA de threat intelligence capable d''analyser et correler les menaces en temps reel	Voir sur GitHub
LogParser-AI	Analyseur de logs propulse par l''intelligence artificielle pour la detection d''anomalies	Voir sur GitHub
YaraMemoryScanner	Scanner memoire utilisant des regles YARA pour la detection de malware en temps reel	Voir sur GitHub
SysmonEventCorrelator	Correlateur d''evenements Sysmon exploitant l''IA pour identifier les patterns d''attaque	Voir sur GitHub
CyberSec-Assistant-3B	Modele de langage 3B parametres specialise en cybersecurite offensive et defensive	Voir sur HuggingFace
CyberSec Leaderboard	Classement des modeles d''IA evalues sur des benchmarks de cybersecurite	Voir sur HuggingFace

Tous ces outils sont disponibles en open source sur notre profil GitHub et nos modeles d''IA sur notre espace HuggingFace. N''hesitez pas a contribuer et a signaler les issues.

Questions Fréquentes

Quelles sont les principales menaces de l'IA offensive en cybersécurité ?
Les principales menaces de l'IA offensive se déclinent en quatre catégories majeures. Premièrement, le phishing automatisé et personnalisé par LLM, qui produit des emails de spear-phishing d'une qualité linguistique indiscernable d'un email légitime, avec un taux de succès

supérieur de 30% aux campagnes traditionnelles. Deuxièmement, les deepfakes audio et vidéo permettant l'usurpation d'identité en temps réel pour la fraude au dirigeant et le contournement des systèmes d'authentification biométrique. Troisièmement, la génération de malwares polymorphes par des outils comme WormGPT et FraudGPT, capables de réécrire automatiquement le code malveillant pour échapper à la détection par signatures. Quatrièmement, l'OSINT automatisé et la reconnaissance augmentée par IA, qui permet une cartographie exhaustive de la surface d'attaque d'une organisation en une fraction du temps traditionnel. L'automatisation complète de la kill chain par des agents IA autonomes constitue une menace émergente particulièrement préoccupante.

Comment l'IA améliore-t-elle la détection des cybermenaces ?
L'IA améliore la détection des cybermenaces à travers plusieurs approches complémentaires. La détection d'anomalies par apprentissage non supervisé (autoencodeurs, modèles de graphes) permet d'identifier des comportements réseau et endpoint déviants sans nécessiter de signatures prédéfinies, détectant potentiellement des menaces zero-day. La détection de malwares par deep learning (CNN sur binaires bruts, analyse comportementale par LSTM/Transformers) atteint des taux de détection supérieurs à 95% sur les datasets de référence. L'analyse comportementale UEBA (User and Entity Behavior Analytics) construit des profils comportementaux individuels pour détecter les compromissions de comptes et les insider threats. Le threat hunting assisté par LLM automatise la génération d'hypothèses de chasse et la corrélation d'indicateurs. Les plateformes SOAR augmentées par IA réduisent le MTTD de 30 à 60% grâce au triage intelligent et à l'enrichissement automatique des alertes.

Qu'est-ce que le OWASP Top 10 LLM et pourquoi est-il important ?
Le OWASP Top 10 pour les applications LLM est un référentiel de sécurité publié par l'Open Web Application Security Project identifiant les dix vulnérabilités les plus critiques spécifiques aux applications intégrant des grands modèles de langage. Il est essentiel car les LLM introduisent des classes de vulnérabilités inédites que les cadres de sécurité traditionnels ne couvrent pas. Les dix vulnérabilités identifiées sont : Prompt Injection (la plus critique, permettant le détournement du modèle), Insecure Output Handling, Training Data Poisoning, Model Denial of Service, Supply Chain Vulnerabilities, Sensitive Information Disclosure (risque de fuite de données d'entraînement ou de données internes via RAG), Insecure Plugin Design, Excessive Agency, Overreliance et Model Theft. Toute organisation déployant des LLM en production devrait conduire une évaluation de sécurité basée sur ce référentiel et implémenter les mesures de mitigation recommandées pour chaque catégorie de vulnérabilité.

Comment se protéger contre les attaques adversariales sur les modèles ML ?
La protection contre les attaques adversariales nécessite une approche multi-couches. Contre les attaques par évasion (FGSM, PGD, C&W), l'adversarial training intègre des exemples adversariaux dans les données d'entraînement pour renforcer la robustesse du modèle, bien que cela implique un compromis robustesse-précision. La distillation défensive et le lissage d'entrée (input smoothing) offrent des protections complémentaires. Contre le data poisoning et les backdoors, les techniques de Neural Cleanse, d'Activation Clustering et de pruning neuronal permettent de détecter et supprimer les comportements malveillants injectés. Contre l'extraction de modèle, le watermarking, le rate limiting des API et la détection de requêtes de distribution anormale limitent les attaques. Le differential privacy (DP-SGD) protège la

confidentialité des données d'entraînement contre les attaques d'inversion et de membership inference. Le framework MITRE ATLAS fournit une taxonomie de référence pour planifier ces défenses de manière systématique.

Que dit l'EU AI Act sur l'utilisation de l'IA en cybersécurité ?
L'EU AI Act, entré en vigueur progressivement depuis 2025, classifie les systèmes d'IA selon quatre niveaux de risque : inacceptable (interdit), haut risque, risque limité et risque minimal. Les systèmes d'IA déployés en cybersécurité dans les secteurs critiques couverts par NIS2 (énergie, transport, santé, finance, administrations) sont susceptibles d'être classés à haut risque. Ces systèmes sont soumis à des obligations strictes : qualité des données d'entraînement (article 10), documentation technique exhaustive (article 11), journalisation des décisions (article 12), transparence envers les utilisateurs (article 13), contrôle humain effectif (article 14) et exigences de précision, robustesse et cybersécurité (article 15). Le non-respect de ces obligations peut entraîner des sanctions allant jusqu'à 35 millions d'euros ou 7% du chiffre d'affaires annuel mondial. Les obligations spécifiques aux LLM et systèmes d'IA à usage général (GPAI) s'appliquent depuis août 2025.

Quels outils IA sont recommandés pour un SOC moderne ?
Un SOC moderne devrait intégrer plusieurs catégories d'outils IA complémentaires. Pour la détection, les solutions EDR/XDR intégrant du ML comme CrowdStrike Falcon, SentinelOne et Microsoft Defender for Endpoint offrent une détection comportementale avancée sur les endpoints. Pour la corrélation et l'analyse, les SIEM augmentés par IA comme Splunk (avec Splunk AI), Microsoft Sentinel (avec Copilot for Security) et Google Chronicle permettent l'analyse de logs en langage naturel et la détection d'anomalies à grande échelle. Pour l'orchestration et la réponse, les plateformes SOAR comme Palo Alto Cortex XSIAM, Swimlane Turbine et IBM QRadar SOAR automatisent le triage, l'enrichissement et la réponse à incident. Pour la threat intelligence, les plateformes intégrant le NLP comme Recorded Future, Mandiant et ThreatConnect automatisent la collecte et l'analyse des renseignements sur les menaces. Le choix des outils doit s'aligner avec l'architecture existante, les compétences de l'équipe et le budget disponible.

Comment prévenir les prompt injections dans les applications LLM d'entreprise ?
La prévention des prompt injections requiert une stratégie de défense en profondeur car aucune technique unique ne garantit une protection complète. Au niveau de la couche d'entrée, implémenter une validation stricte des prompts avec des classificateurs ML entraînés à détecter les patterns d'injection, utiliser des séparateurs structurels clairs entre instructions système et données utilisateur, et limiter la taille et le format des entrées. Au niveau du modèle, utiliser le prompt engineering défensif avec des instructions explicites de refus, implémenter l'instruction hierarchy (les instructions système ont une priorité hiérarchique stricte sur les données utilisateur), et fine-tuner le modèle pour reconnaître et refuser les tentatives d'injection. Au niveau de la sortie, filtrer les réponses pour détecter les fuites de données sensibles, valider les actions avant exécution, et journaliser toutes les interactions pour la détection d'anomalies. Contre l'injection indirecte, sanitiser toutes les données externes avant injection dans le contexte, implémenter des contrôles de confiance différenciés selon la source des données, et limiter les capacités d'action du modèle au strict minimum nécessaire.

Quel est le rôle du MITRE ATLAS dans la sécurité des systèmes d'IA ?

MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems) est le pendant du célèbre MITRE ATT&CK spécifiquement conçu pour les menaces ciblant les systèmes d'intelligence artificielle et de machine learning. Son rôle principal est de fournir une taxonomie structurée des tactiques, techniques et procédures (TTPs) adversariales spécifiques à l'IA, organisée en quatorze tactiques couvrant l'ensemble du cycle de vie des systèmes ML : de la reconnaissance de l'infrastructure ML jusqu'à l'impact sur les prédictions, en passant par le data poisoning, l'extraction de modèle et l'évasion adversariale. Chaque technique est documentée avec des études de cas réelles, des procédures d'exploitation et des recommandations de mitigation. ATLAS est utilisé par les organisations pour conduire des threat modeling systématiques de leurs déploiements d'IA, évaluer leur posture de sécurité par rapport aux menaces connues, planifier des exercices de red teaming ciblés, et prioriser les investissements en sécurité. Il est complémentaire de l'OWASP Top 10 LLM (plus focalisé sur les vulnérabilités applicatives) et du NIST AI RMF (plus focalisé sur la gouvernance).

Comment l'intelligence artificielle transforme-t-elle la cybersecurite ?

L'intelligence artificielle transforme la cybersecurite sur deux fronts. Offensivement, elle permet l'automatisation de la reconnaissance, la generation de phishing ultra-cible, la creation de deepfakes convaincants, le contournement des systemes de detection et l'optimisation des chaines d'exploitation. Defensivement, l'IA ameliore la detection d'anomalies, l'analyse comportementale des utilisateurs (UEBA), la classification automatisee des menaces, la reponse automatisee aux incidents et l'analyse de malwares. Cette course aux armements IA vs IA definit le paysage de la cybersecurite moderne.

Quelles sont les principales attaques adversariales contre les modeles IA ?

Les principales attaques adversariales incluent les attaques par evasion (modification subtile des entrees pour tromper le modele en production), l'empoisonnement de donnees (injection de donnees malveillantes dans le jeu d'entrainement), l'extraction de modele (vol de la propriete intellectuelle par requetes systematiques), l'inference de membres (determination si une donnee specifique fait partie du jeu d'entrainement), et les attaques par injection de prompt contre les LLM. Chaque type d'attaque cible une phase differente du cycle de vie du modele.

Comment les grands modeles de langage sont-ils utilises en cybersecurite offensive ?

Les LLM sont utilises offensivement pour generer du code malveillant polymorphe, creer des emails de phishing linguistiquement parfaits et personnalises, automatiser l'analyse de code source pour identifier des vulnerabilites, generer des payloads d'exploitation adaptatifs, et conduire des operations de desinformation a grande echelle. Ils permettent egalement de contourner les filtres de securite par des techniques de jailbreak et d'analyser automatiquement la documentation technique d'une cible pour identifier des vecteurs d'attaque potentiels.

Quel est l'impact du EU AI Act sur la cybersecurite ?

Le EU AI Act classe les systemes d'IA par niveau de risque et impose des exigences proportionnees. Pour la cybersecurite, les systemes de detection de menaces et de surveillance sont consideres comme a haut risque, necessitant une documentation technique complete, des evaluations de conformite, une gestion des biais, une supervision humaine et une transparence vis-a-vis des utilisateurs. Les outils d'IA offensive utilises en recherche doivent respecter des cadres ethiques stricts. Le non-respect peut entrainer des amendes pouvant atteindre 35 millions d'euros ou 7% du chiffre d'affaires mondial.

Pour approfondir, consultez les ressources de NIST Cybersecurity et de NVD (National Vulnerability Database).

Sources et références : ANSSI · CERT-FR

Comment detecter les contenus generes par intelligence artificielle ?

La detection de contenus generes par IA repose sur plusieurs approches : l'analyse statistique des distributions de tokens (les LLM produisent des distributions caracteristiques), la detection de watermarks numeriques integres lors de la generation, l'analyse des metadonnees et de la provenance du contenu, les classificateurs entraines specifiquement sur des contenus IA vs

humains, et l'analyse forensique des images generees par diffusion. Cependant, les techniques de detection evoluent constamment face aux ameliorations des modeles generatifs, rendant cette course technologique permanente.

Ayi NEDJIMI Consultants — Expert cybersécurité offensive & intelligence artificielle

ayinedjimi-consultants.fr · ayi@ayinedjimi-consultants.fr