L'usage offensif de l'Intelligence Artificielle par les cyberattaquants

Par Lucien CHAYA PODEUR expert XMCO

Introduction

En novembre 2022, l'introduction de ChatGPT par la firme OpenAI a permis une démocratisation sans précédent des technologies de Natural Language Processing (NLP) auprès du grand public et, par voie de conséquence, des cybercriminels qui les ont détournés à des fins malveillantes.
L'intégration de l'IA dans les modes opératoires des cybercriminels pourrait provoquer une inflation du volume des intrusions, facilitées par des outils d'automatisation et d'optimisation pilotés par IA.

Les organisations ciblées doivent désormais composer avec des menaces polymorphes, adaptatives, et en constante évolution, bien qu'étant encore limitées par les contraintes techniques de développement des modèles.

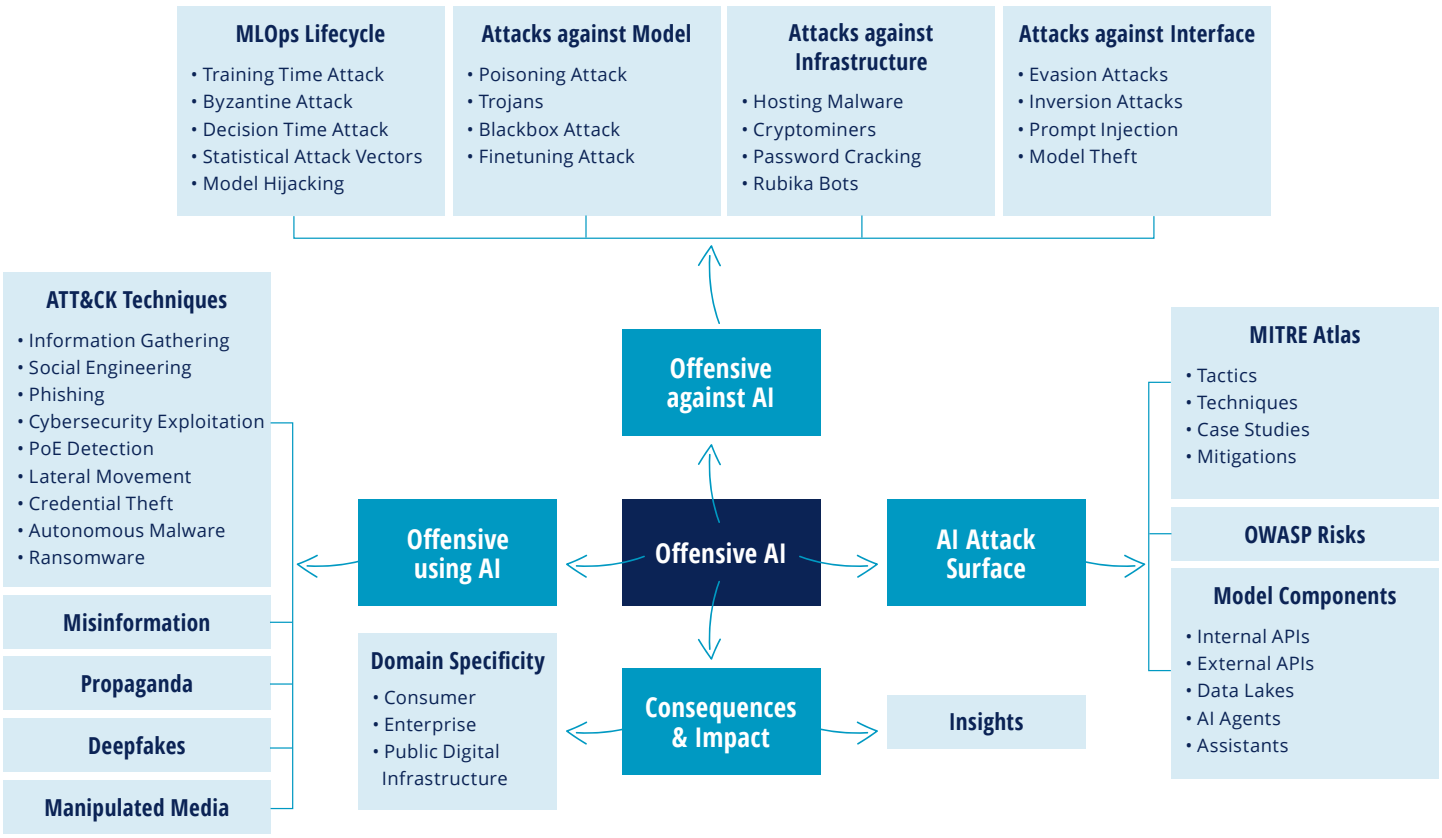
En partenariat avec OpenAI, les chercheurs de Microsoft ont annoncé le 14 février 2024 que des groupes APT liés à la Russie, à l'Iran, à la Corée du Nord et à la Chine avaient recours aux LLM (Large Language Model) dans le cadre de leurs opérations[1]. Ces groupes employaient les outils d'OpenAI pour collecter des informations en sources ouvertes, élaborer des campagnes de phishing, traduire des documents d'intérêt, détecter des erreurs de programmation et automatiser des tâches courantes.

Comprendre les leviers de cette transformation suppose d'analyser la façon dont sont construits les modèles d'IA les plus utilisés dans ce contexte, en particulier les LLM. Leur création s'articule en plusieurs étapes clefs : la collecte massive de données textuelles, leur prétraitement pour éliminer les biais et les erreurs, l'entraînement du modèle via des algorithmes d'apprentissage profond, puis le fine-tuning qui permet d'adapter le modèle à des tâches précises. La phase de déploiement s'accompagne finalement d'une supervision continue couplée à des mécanismes de retour d'expérience pensés pour améliorer la robustesse et la pertinence du modèle face à de nouveaux usages ou à des tentatives de contournement malveillantes. Malgré ces dispositifs de protection, les LLM demeurent exposés à des attaques, qu'il s'agisse d'injection de prompts, d'empoisonnement des données ou d'exploitation de failles dans leur processus de développement, ces dernières pouvant porter atteinte à leur fiabilité et à leur sécurité, limitant par conséquent leur adoption généralisée.

L'analyse développée dans cet article s'appuie sur le modèle de la kill chain cyber, un cadre méthodologique permettant de décomposer les différentes étapes d'une attaque informatique, depuis la phase de reconnaissance jusqu'à l'exfiltration des données ou la compromission prolongée du système ciblé[2].

L'usage offensif de l'intelligence artificielle bouleverse aujourd'hui chacune de ces étapes : automatisation de la phase de renseignement, production de campagnes de phishing sur mesure, développement de malware polymorphes, ou encore utilisation de deepfakes pour tromper les contrôles humains.

Cette dynamique se manifeste dès la phase de reconnaissance : traditionnellement constituée de recherches manuelles chronophages, elle connaît désormais une mutation profonde grâce à l'apport de l'IA.



Vue d'ensemble de l'IA appliquée aux méthodes offensives. Source : Arxiv[3]

Glossaire

Adversarial Attack
Technique visant à tromper un modèle d'Intelligence Artificielle (IA) ou de Machine Learning (ML) en lui fournissant des entrées spécifiquement conçues pour provoquer des erreurs de classification ou des comportements inattendus. Ces attaques exploitent les vulnérabilités des modèles, notamment des réseaux neuronaux profonds, en modifiant subtilement les données d'entrée pour induire des prédictions erronées.

Adversarial Learning
Domaine de la sécurité de l'IA qui étudie la résistance des modèles face aux attaques adversariales. Il s'agit de comprendre comment des acteurs malveillants peuvent manipuler les données d'entrée ou d'entraînement pour biaiser ou tromper un modèle, et de développer des contre-mesures pour renforcer la robustesse des systèmes d'IA.

Classification
Dans le contexte de l'intelligence artificielle (IA), la classification désigne une méthode d'apprentissage automatique consistant à attribuer une catégorie ou une classe à une donnée d'entrée selon ses caractéristiques.

Deepfake
Contenu audio, vidéo ou image généré par IA, imitant de manière réaliste une personne réelle. Les deepfakes sont utilisés pour des fraudes, des manipulations ou des attaques d'ingénierie sociale sophistiquées.

Empoisonnement de modèles *(Data Poisoning/Model Poisoning)*
Type d'attaque adversariale visant à insérer des données corrompues ou biaisées dans le jeu d'entraînement d'un modèle d'IA ou de ML. L'objectif est de manipuler le comportement du modèle, en induisant des prédictions erronées ou des failles de sécurité.

Fine-tuning
Processus d'ajustement d'un modèle d'IA pré-entraîné sur un large corpus, pour le spécialiser sur une tâche ou un domaine particulier, en utilisant un jeu de données plus restreint et spécifique.

GAN *(Generative Adversarial Network / Réseau Génératif Antagoniste)*
Architecture de deep learning composée de deux réseaux neuronaux antagonistes : un générateur, qui crée de nouvelles données et un discriminateur, qui tente de distinguer les données générées des données réelles. Les GAN sont utilisés pour produire des images, des sons ou du texte réalistes et sont à la base de nombreuses applications de deepfake.

GenAI *(Intelligence Artificielle Générative)*
Ensemble des technologies d'IA capables de générer de nouveaux contenus (texte, images, audio, vidéo) de manière autonome, à partir de modèles entraînés sur de vastes ensembles de données. GenAI inclut les LLM (Large Language Models), les GAN et d'autres architectures génératives. Il est utilisé aussi bien pour des usages légitimes que malveillants.

Groupe APT *(Advanced Persistent Threat)*
Ensemble structuré d'attaquants, souvent soutenus par des États ou des organisations criminelles, menant des opérations cyber offensives avancées, ciblées et persistantes.

Kill Chain
Modèle décrivant les étapes successives d'une cyberattaque, de la reconnaissance initiale à l'impact final. Il permet d'analyser et de structurer les différentes phases d'une intrusion pour mieux comprendre le mode opératoire des attaquants.

Large Language Model *(LLM)*
Modèle d'IA de grande taille, entraîné sur d'immenses volumes de textes, capable de comprendre, générer et manipuler du langage naturel. Les LLM sont utilisés pour la génération de texte, l'analyse sémantique, la traduction, etc.

Malware polymorphe
Logiciel malveillant qui modifie son code ou sa structure à chaque infection, afin d'échapper aux systèmes de détection basés sur des signatures statiques.

MLOps *(Machine Learning Operations)*
Ensemble des pratiques visant à industrialiser, déployer, superviser et maintenir les modèles de machine learning en production, tout en assurant leur sécurité et leur robustesse.

NLP *(Natural Language Processing / Traitement Automatisé du Langage Naturel)*
Branche de l'IA qui permet aux ordinateurs d'analyser, de comprendre et de générer le langage humain. Le NLP est à la base des assistants vocaux, des traducteurs automatiques et des LLM.

Prompt
Instruction ou question formulée par un utilisateur pour guider une intelligence artificielle dans la génération d'une réponse ou d'un contenu spécifique.

Une phase de reconnaissance automatisée

La collecte d'informations représente une phase préliminaire au cours de laquelle les attaquants rassemblent des renseignements sur leurs cibles pour comprendre l'environnement, identifier les vulnérabilités potentielles et adapter leurs stratégies d'attaque.

Cette phase comprend deux approches principales : la collecte active, impliquant une interaction directe avec les systèmes cibles via des outils comme Nmap pour les scans de réseau, et la collecte passive, plus discrète, reposant sur l'analyse de données publiquement accessibles (OSINT) comme les bases WHOIS, les moteurs de recherche et les médias sociaux.

Collecte de données et cartographie des données

L'IA a introduit de nouvelles approches dans la phase de reconnaissance et de collecte de données, en facilitant l'automatisation et l'analyse des informations. Jusqu'à l'intégration de systèmes intelligents au sein du processus, cette étape reposait principalement sur des méthodes manuelles ou semi-automatisées[4]. L'IA offre désormais des capacités avancées de traitement des données collectées, permettant une cartographie précise et une visualisation globale de l'entreprise ciblée, fournissant ainsi aux auditeurs de sécurité ainsi qu'aux potentiels attaquants une vue d'ensemble structurée de leur cible.

Les outils légitimes, enrichis par l'apport de modèles d'IA, offrent aujourd'hui des fonctionnalités avancées susceptibles d'être exploitées aussi bien par des professionnels de la cybersécurité que dans le cadre d'un usage détourné, par des acteurs de la menace. C'est le cas de Nmap, initialement conçu comme un scanner réseau pour administrateurs système, permettant de découvrir les hôtes et services sur un réseau informatique[5]. Le projet nmap.ai améliore les fonctionnalités classiques de Nmap en y intégrant une couche d'IA analysant automatiquement les résultats techniques du scan[6]. Grâce à l'utilisation d'un modèle GPT-3.5, nmap.ai transforme les sorties brutes de Nmap en points clefs simplifiés et fournit des recommandations de sécurité qui, détournées de leur usage initial, peuvent être exploitées par les attaquants.

L'émergence de bots IA dédiés au scraping de plateformes comme LinkedIn offre de nouvelles perspectives dont certains groupes cybercriminels pourraient bénéficier. Une étude réalisée par l'Université de Stanford en 2024 démontre que l'utilisation de scrapers ayant recours à l'IA permet aujourd'hui d'extraire et de structurer les informations de plus de 100 000 profils LinkedIn en moins de 24 heures, là où une opération manuelle équivalente aurait nécessité un travail humain autrement plus conséquent[7].

Les données collectées permettent ensuite aux attaquants de cartographier les structures d'entreprises ciblées et d'identifier des points d'entrée probants, notamment via le croisement de ces informations avec des bases de données issues de fuites ou des moteurs de recherche spécialisés dans l'indexation de ressources exposées sur internet (Shodan, Censys). Cette approche permet aux attaquants de dresser un portrait détaillé de leurs cibles, facilitant la préparation d'attaques sur mesure et la sélection des vecteurs d'intrusion les plus pertinents.

Reconnaissance en vulnérabilité

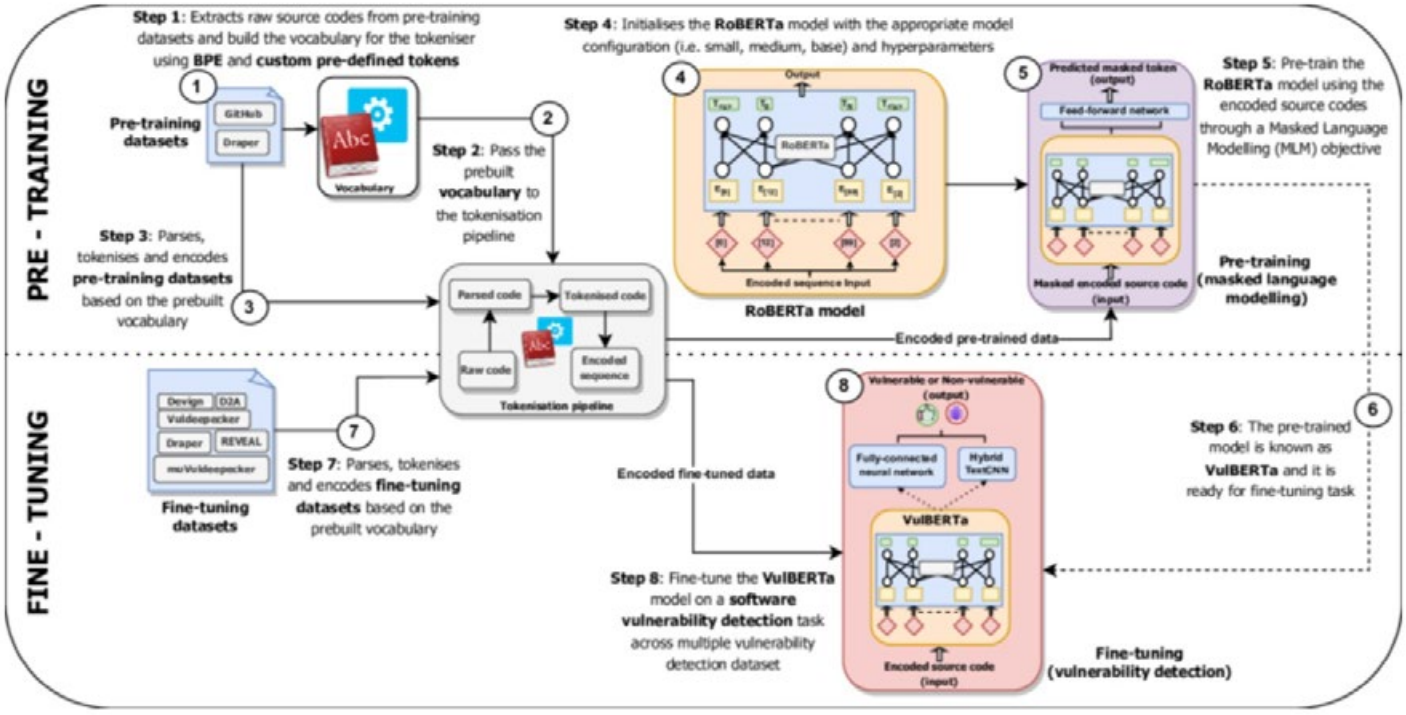
L'intégration de l'IA démultiplie également l'efficacité des outils utilisés dans la détection des vulnérabilités affectant les systèmes ciblés. De manière similaire, cette étape reposait traditionnellement sur des analyses manuelles

ou semi-automatisées via des scripts.
Désormais, des systèmes spécifiquement entraînés à l'analyse des descriptions de vulnérabilités informatiques tels que VulBERTA sont capables de traiter des milliers de documentations techniques et d'identifier sans intervention humaine les brèches critiques propres à chaque organisation, tout en recommandant des stratégies d'exploitation adaptées à l'infrastructure technologique de la victime[8]. Utilisé en parallèle d'outils réalisant des scans massifs, VulBERTA permet de croiser les données collectées avec une analyse sémantique approfondie du code source ciblé. Selon l'étude publiée par l'ETH Zurich à l'origine du logiciel, VulBERTA a permis d'augmenter de 35 % la rapidité de détection des vulnérabilités exploitables par rapport aux méthodes traditionnelles[9].

Au-delà de la simple identification, l'IA corrèle les vulnérabilités repérées avec les renseignements collectés durant la phase de reconnaissance afin de cibler en priorité les failles réellement présentes dans l'environnement technique de la victime, puis interroge automatiquement les bases de données de vulnérabilités pour identifier les failles non corrigées correspondantes.

De façon comparable, Burp Suite tire parti de l'intelligence artificielle à travers Burp AI pour aller au-delà de la simple détection, en proposant une analyse approfondie et contextualisée des vulnérabilités[11]. Burp est un outil modulaire permettant de réaliser des tests manuels ou automatisés aidant les analystes en sécurité aussi bien que les cybercriminels à identifier des vulnérabilités sur les applications web. Désormais, grâce à l'intégration de l'IA, Burp ne se limite plus à l'identification automatisée des failles : l'outil est capable d'analyser en profondeur les vulnérabilités détectées, en générant des preuves d'exploitation concrètes et en proposant des stratégies d'attaque adaptées à la configuration spécifique de l'application cible.
Utilisé en synergie avec des scans massifs, Burp AI croise les résultats du scanner avec une analyse contextuelle et sémantique des flux applicatifs, permettant de valider l'impact réel des vulnérabilités dans l'environnement testé.

Système d'entraînement de VulBERTA
Source : ResearchGate[10]

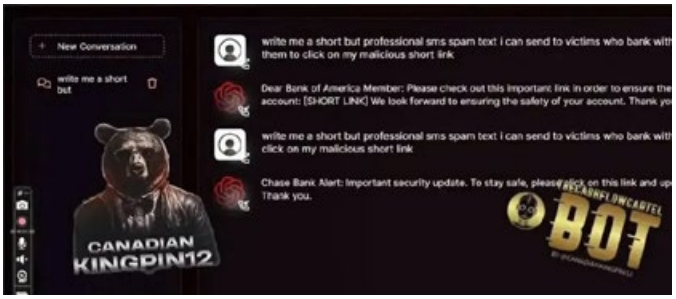


Accès initial et social engineering

Deux tendances majeures se dégagent de l'implémentation de l'IA au sein des techniques d'ingénierie sociale : le phishing hyper personnalisé et l'utilisation de deepfakes audio/vidéo désormais accessibles à moindre coût. ■

Phishing

Les modèles de traitement du langage naturel tels que ChatGPT, ou l'un de ses dérivés malveillants tels que FraudGPT, sont désormais employés pour générer des messages personnalisés, adaptés au contexte et au profil psychologique des cibles[12]. Les indices traditionnels utilisés pour détecter les emails de phishing tels que les erreurs grammaticales et les tons inadéquats employés par les attaquants sont aujourd'hui des écueils largement atténués par l'intégration de l'IA dans le processus de rédaction. Le fait que ces emails se déclinent en plusieurs variantes permet également de mettre à mal les règles de détection anti-phishing n'incluant pas de mesures anti-IA[13]. Une étude menée par l'Université de San Antonio révèle que les taux d'efficacité des emails générés par IA sont très proches, voire supérieurs à ceux des emails écrits par des humains et reproductibles à volonté[14]. Une seconde étude démontre qu'au premier trimestre 2025, 82,6% des emails de phishing analysés utilisaient l'IA, avec une augmentation de 17,3% du volume global par rapport au semestre précédent[15].



Génération de SMS de phishing à l'aide de FraudGPT
Source : Netenrich[12]

Les campagnes de phishing s'appuient fréquemment sur des événements d'actualité dans le but de renforcer l'illusion de légitimité. Dans le cadre des Jeux Olympiques s'étant tenus à Paris au cours de l'été 2024, les chercheurs de Trend Micro ont analysé une campagne incitant les potentielles victimes à investir dans des cryptomonnaies frauduleuses[16].

Les attaquants ont bâti cette campagne autour de sites web frauduleux, agrémentés d'images produites par intelligence artificielle pour accroître leur crédibilité. Ce recours à l'ingénierie sociale et à la contextualisation des attaques se retrouve également dans d'autres formes de fraudes.

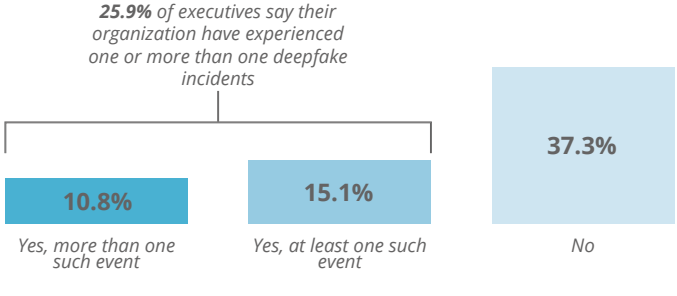
Le compte X (Twitter) vérifié d'Olympics_Solana diffusant son site frauduleux. Source : Trend Micro[16]

Les attaques de type *Business Email Compromise* (BEC) illustrent parfaitement cette évolution. Dans ce type de fraude, les cybercriminels usurpent l'identité d'un dirigeant ou d'un partenaire commercial pour manipuler un employé en créant un sentiment d'urgence ou de confidentialité pour le pousser à procéder à des virements frauduleux ou à l'envoi de données sensibles vers des comptes contrôlés par l'attaquant. Ces mails, générés automatiquement grâce à des modèles d'IA générative en fonction des informations collectées lors de la phase de reconnaissance, constituent aujourd'hui l'un des scénarios de compromission présentant le plus fort taux de réussite[14]. Selon les données du FBI, cette menace a causé des pertes financières colossales atteignant 55,5 milliards de dollars entre octobre 2013 et décembre 2023, avec plus de 305 033 cas signalés[17]. L'essor de l'IA dans la création de campagnes de phishing a contribué à une augmentation rapide du

nombre d'attaques, celles-ci devenant de plus en plus crédibles et sophistiquées[18].
L'intelligence artificielle facilite désormais l'orchestration d'attaques multicanales sophistiquées combinant phishing, vishing (voice phishing) et deepfake. Cette méthode d'attaque particulièrement efficace associe l'envoi initial d'un email frauduleux suivi par un contact téléphonique lors duquel les attaquants emploient des voix générées artificiellement pour manipuler leurs victimes. Les chercheurs de Sophos ont observé cette méthode lors de 15 incidents de sécurité entre novembre 2024 et la mi-janvier 2025. Les opérateurs du ransomware 3AM ont d'abord diffusé massivement des emails sur une courte période, puis contacté les utilisateurs ciblés par téléphone pour les informer d'un incident de sécurité et les accompagner dans la remédiation, facilitant ainsi la prise de contrôle des systèmes visés[19].
Cette convergence technologique permet aux attaquants de renforcer la crédibilité de leurs tentatives d'ingénierie sociale en créant une cohérence entre les différents vecteurs d'attaque. L'IA coordonne ces différentes phases, adaptant le discours téléphonique en fonction des réactions au message initial, créant ainsi une expérience d'ingénierie sociale fluide et convaincante.

Deepfakes

Ces technologies permettent de créer des imitations quasi parfaites de voix et de visages, rendant les attaques d'ingénierie sociale considérablement plus convaincantes[20].
En février 2024, l'entreprise d'ingénierie britannique Arup a perdu plus de 25 millions de dollars suite à une fraude sophistiquée impliquant l'usage d'un deepfake au cours d'une vidéoconférence au sein de laquelle le directeur financier et d'autres collègues étaient générés par IA[21]. Initialement méfiant face à un message demandant une transaction secrète, l'employé a finalement été convaincu par le réalisme des participants virtuels à l'appel vidéo. Loin d'être isolé, cet incident s'inscrit dans un contexte global marqué par la multiplication des attaques ayant recours à l'IA pour contourner les dispositifs de sécurité, notamment à travers la compromission de données biométriques destinées à alimenter des deepfakes.
Le 15 février 2024, Group-IB a par exemple mis au jour un nouveau Trojan bancaire baptisé GoldPickaxe.iOS, ciblant les utilisateurs iOS et étant capable de collecter les données de reconnaissance faciale de ses victimes[22]. Les acteurs de la menace exploitent ensuite ces données biométriques pour générer des deepfakes via des outils de face-swapping leur permettant de remplacer leur visage par celui des victimes et accéder à des informations sensibles ou des comptes bancaires[23].
Cette capacité à générer des deepfakes à partir de données biométriques volées ouvre désormais la voie à des attaques encore plus sophistiquées, notamment grâce à l'émergence de technologies de deepfake en temps réel.
En l'état, la réalisation d'un deepfake live est encore réservée aux acteurs sophistiqués disposant d'une puissance de calcul conséquente. Des outils tels que *"Deep-Live-Cam"* permettent aujourd'hui d'en réaliser mais sont encore limités par leurs performances. Les caractéristiques physiques poussées, à l'image des lunettes ou de la pilosité compliquent significativement la réalisation[24]. À ce jour, l'attaque la plus accessible impliquant un deepfake demeure la réalisation d'une vidéo préalable ou d'un message audio pouvant être envoyé à la victime. Il est néanmoins plausible que les technologies de deepfake live se développement et se démocratisent dans les prochaines années, dans le sillon des progrès réalisés par l'IA.
La menace demeure significative : d'après un sondage réalisé par Deloitte, 25,9 % des dirigeants interrogés ont révélé que leur organisation avait connu un ou plusieurs incidents impliquant des deepfake ciblant des données financières et comptables au cours des 12 mois précédents[25].



Don't know / not applicable: 36.8%

En réponse à la question : « During the past 12 months, did your organization experience any deepfake incidents targeting financial andaccounting data? ». Source : Deloitte[25]

Exploitation, déploiement, obfuscation

Génération de code malveillant par IA

L'intelligence artificielle générative (GenAI) permet de générer des solutions ou des contenus inédits, pour autant l'accessibilité des modèles génératifs n'a pas produit d'explosion du nombre de nouveaux malware dans la nature[26]. Les systèmes actuels de GenAI ne disposent pas des capacités spécifiques pour créer de manière indépendante des malware opérationnels, et requièrent de fait une intervention humaine afin de corriger et diriger le processus de création[27]. L'efficacité du modèle dépendant de ses données d'entraînement, la qualité de la génération en pâtit car les exemples de malware sophistiqués sont rarement accessibles publiquement.

En outre, pour qu'un système d'intelligence artificielle générative puisse concevoir un malware sophistiqué fonctionnel, il doit non seulement être apte à produire du code robuste mais également disposer d'informations relatives à des failles exploitables concernant les systèmes ciblés. Les exigences poussées liées à la génération de malware limitent donc son utilisation à un cercle restreint d'acteurs possédant ces capacités et informations[28]. Les cybercriminels utilisent néanmoins la GenAI pour créer des malware ou des scripts simples, améliorer les compétences des malware existants, ou en créer des variantes.

```
$ie = New-Object -ComObject "InternetExplorer.Application"
$ie.visible = $false
$ie.navigate("https://www.example.com/login")
while ($ie.Busy -eq $true) { Start-Sleep -Milliseconds 100 }
$usernameField = $ie.Document.getElementByID("username")
$usernameField.value = "username"
$passwordField = $ie.Document.getElementByID("password")
$passwordField.value = "password"
$submitButton = $ie.Document.getElementByID("submit")
$submitButton.click()
Start-Sleep -Seconds 5
$cookie = $ie.Document.cookie
$cookie | Out-File -FilePath "C:\Path\To\WebSessionCookie.txt"
```

Prompt demandant à ChatGPT de créer du code à partir d'une technique MITRE ATT&CK. Source : Trend Micro[26]

Des exemples de génération de malware dans la nature ont déjà été observés par des chercheurs. HP Wolf Security a mis au jour une campagne ciblant des utilisateurs francophones au cours de laquelle des JavaScript et VBScript malveillant ont été rédigés à l'aide d'une IA générative[29]. La structure, la présence de commentaires détaillant chaque ligne de code, ainsi que le choix de noms de fonctions a permis aux chercheurs de déterminer que de la GenAI avait été utilisée par les cybercriminels afin de développer ces scripts. Cette campagne visait à infecter les victimes avec AsyncRAT, un infostealer capable d'enregistrer les frappes du clavier et l'écran de l'utilisateur. Si ce cas met en lumière la manière dont l'IA générative abaisse les barrières techniques lors de la conception d'attaques, il convient de souligner que cette technologie demeure à un stade précoce de développement, et que ses applications futures pourraient démultiplier l'ampleur et la complexité des menaces.

OpenAI a annoncé en mai 2025 mettre à disposition du public son agent d'ingénierie logicielle Codex, capable d'écrire, comprendre et corriger du code informatique. Entraîné via un apprentissage par renforcement sur des tâches de programmation réelles afin de produire du code conforme aux standards humains, cet outil pourrait être détourné de ses usages légaux afin de servir les intérêts de cybercriminels. Comme observé avec les versions jailbreakées de ChatGPT, les restrictions de sécurité implémentées par OpenAI pourraient très probablement être contournées et les capacités de Codex seraient alors utilisées à des fins malveillantes[30].

Le Center for Emerging Technology and Security a publié en juillet 2024 une analyse des potentiels usages de l'IA générative dans le cadre de la génération de malware [28]. Selon les chercheurs, l'IA générative pourrait permettre à des agents malveillants de modifier leur code à l'exécution, voire de se réécrire entièrement afin d'échapper à la détection. Ces agents seraient également capables de rédiger eux-mêmes des payloads ou de créer de nouveaux outils pour surmonter des obstacles inédits, adaptant ainsi leurs tactiques en temps réel et opérant de manière autonome, avec un besoin réduit de supervision humaine. La coopération entre plusieurs agents offrirait une persistance renforcée, chaque entité

apprenant et s'adaptant continuellement à son environnement.
Enfin, leur capacité à raisonner sur leur contexte et à ajuster leurs communications pour se fondre dans le trafic légitime conférerait à ces malware une furtivité et une persistance sans précédent.

Techniques d'obfuscation, malware polymorphiques

L'IA offre la possibilité de multiplier les variantes de malware et de scripts, compliquant ainsi la tâche des systèmes de détection fondés sur les règles YARA. Recorded Future l'ont démontré en adaptant le code de l'infostealer STEELHOOK, utilisé par le groupe d'attaquants APT28 attribué à la Russie, pour contourner les règles YARA en étudiant précisément leurs modes de détection[31].
Une équipe de chercheurs de l'Indian Institute of Technology Madras a conduit une analyse approfondie de l'usage de l'IA dans le cadre d'actions cyber offensives[32]. Le papier analyse comment cette nouvelle technologie pourrait obliger la cybersécurité à repenser son arsenal de détection, en distinguant trois évolutions majeures rattachées à des phases distinctes de la killchain :

• **L'accès initial** *(Point of Entry)* : d'un point de vue défensif, cette phase initiale est essentielle pour identifier et bloquer les menaces dès leur apparition. L'article présente Deeplocker, développé en tant que proof-of-concept par IBM Research, illustrant l'utilisation de l'IA pour créer des malware hautement furtifs[33].
L'innovation clef réside dans l'intégration d'un réseau de neurones profond (DNN) pour dissimuler une charge malveillante au sein d'une application légitime. La payload est dissimulée grâce à l'IA, son déchiffrement dépendant d'une clef générée dynamiquement par le DNN. Aucune clef n'étant stockée dans le code, la rétro-ingénierie devient inefficace. La payload demeure inactive jusqu'à ce que le malware détecte une combinaison précise d'attributs (biométriques, de configuration). DeepLocker matérialise le risque des attaques IA embarquées, où l'IA sert à la fois de verrou (dissimulation) et de clef (activation ciblée se basant sur de multiples variables), nécessitant une évolution des paradigmes de détection.

• **L'évasion :** cette phase se concentre sur les techniques utilisées par les cyberattaquants pour échapper à la détection au sein d'un réseau après l'avoir pénétré. L'IA, et notamment l'aspect d'apprentissage par renforcement reposant sur le principe d'essai et erreur, permet au malware de modifier sa structure interne afin de contourner la détection statique lors de phases de test contre des antivirus. Des agents IA peuvent interagir dynamiquement avec des échantillons de malware pour appliquer des transformations binaires qui préservent la fonctionnalité du code tout en contournant les systèmes de détection statique. Ces manipulations incluent par exemple l'ajout de caractères ou d'octets aléatoires, ou encore la suppression de signatures, permettant au malware de rester indétecté.

• **Prise de décision lors du déploiement de la charge :** le malware autonome se caractérise par son utilisation de l'IA afin de prendre des décisions éclairées et instantanées, conduisant à des cyber-attaques autonomes[34]. Il est capable d'évaluer l'environnement et prendre des décisions avisées grâce à des instructions dynamiques et d'adapter ses décisions en temps réel en fonction des besoins.

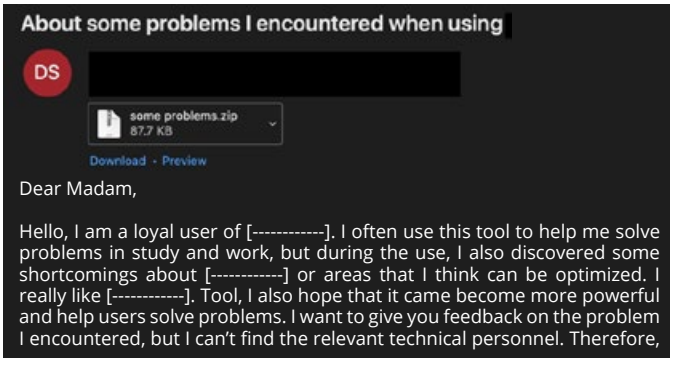
Cette nouvelle technologie de génération permet aux malware d'évoluer vers des formes polymorphes, générées dynamiquement pour échapper aux signatures des antivirus et s'adapter en temps réel à son environnement.
Chaque nouvelle itération d'un ransomware peut par exemple présenter des variations dans son code source, ses chaînes de caractères ou ses méthodes de chiffrement, ce qui rend inefficaces les bases de signatures classiques. Cette automatisation permet aux attaquants de lancer des campagnes massives, tout en réduisant le risque d'être détectés par les outils de sécurité statiques.

De la sécurité des modèles d'IA

Le MLOps, ou Machine Learning Operations, est une fonction clef de l'ingénierie du machine learning englobant tous les processus nécessaires pour mettre les modèles d'IA en production et les maintenir. Le cycle de vie MLOps comporte plusieurs phases essentielles : le développement, la phase de test, le déploiement, la supervision et enfin les retours d'information. Toutes ces phases sont susceptibles d'être victimes d'attaques de différentes natures[35].

Les outils basés sur l'IA exploités comme vecteur d'accès initial

À mesure que l'intelligence artificielle s'implante dans les environnements professionnels, les outils qui en découlent font désormais l'objet d'attaques visant à les compromettre afin de s'introduire dans les systèmes d'information liés. Ces offensives ont pour principal objectif de dérober des informations sensibles, d'autant plus que les systèmes d'IA eux-mêmes présentent désormais des vulnérabilités documentées et activement exploitées, telles que la vulnérabilité référencée CVE-2025-31693 au sein de Drupal AI, permettant de compromettre l'intégrité ou la confidentialité des données manipulées par ces systèmes[36].
Plus encore, le 24 avril 2024, des chercheurs de Cornell Tech, de l'Israel Institute of Technology et d'Intuit ont présenté Morris II, un ver informatique inédit exploitant les services d'intelligence artificielle générative pour se propager et mener des actions malveillantes[39]. Construit autour d'un prompt autoréplicateur, Morris II exploite la capacité des LLM à générer et propager automatiquement des instructions malveillantes, à la manière d'une injection SQL. Ce ver est capable de s'auto-répliquer et de se diffuser dans des écosystèmes d'agents GenAI interconnectés, notamment via des assistants de messagerie dotés d'IA, sans nécessiter d'interaction de la part de l'utilisateur (« zero-click »). Les chercheurs ont démontré que Morris II peut exfiltrer des données sensibles telles que noms, numéros de téléphone, informations bancaires ou numéros de sécurité sociale, et lancer des campagnes de spam, en exploitant les failles des assistants de messagerie alimentés par des modèles comme ChatGPT, Gemini ou LLaVA.
Des vulnérabilités référencées ciblant des outils intervenant dans l'étape de développement de modèles d'IA ont également commencé à faire leur apparition. C'est le cas de la vulnérabilité critique CVE-2025-3248 affectant Langflow, un outil open source populaire pour la création de workflows d'agents IA, permettant un attaquant distant non authentifié d'exécuter du code arbitraire sur le serveur cible[40][41].
Une exploitation réussie pourrait mettre en péril la confidentialité des modèles, des jeux de données, des flux d'agents IA en développement et de toute la chaîne d'approvisionnement logicielle. L'analyse de cette vulnérabilité souligne la nécessité de restreindre l'exposition des outils IA récents et de privilégier leur déploiement en environnement isolé, afin d'éviter tout risque d'empoisonnement de modèles.
Du fait des opportunités présentées par la compromission de la chaîne d'approvisionnement logicielle associée à l'IA, le secteur devient une cible privilégiée pour des cyberattaques cherchant à perturber les processus scientifiques[37]. Le 16 mai 2024, Proofpoint a publié un rapport détaillant la distribution du malware SugarGh0st RAT par des acteurs malveillants visant spécifiquement des organisations américaines engagées dans la recherche et le développement en intelligence artificielle, dont des universités, des entreprises privées et des agences gouvernementales[38].



I can only send these questions to you, hoping that you can help me solve them or provide fedback to relevant personnel. Tha,ks for your help!

Sincerely,
Derrick Sean

Mail frauduleux distribuant SugarGh0st RAT
Source : Proofpoint[38]

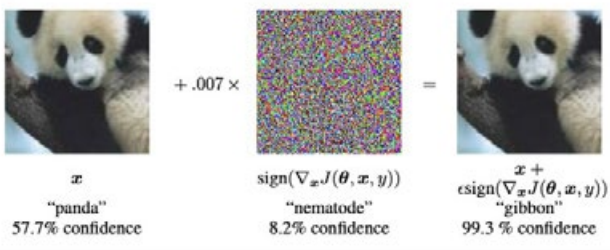
Cette campagne illustre une tendance croissante : les entités impliquées dans la recherche sur l'IA représentent désormais un enjeu majeur pour des opérations de cyberespionnage, les attaquants cherchant à obtenir des renseignements sensibles et non publics sur les avancées technologiques en intelligence artificielle. Elles pourraient en outre permettre à des attaquants d'empoisonner le code et / ou les données sur lesquelles sont entraînés les modèles, afin de compromettre en cascade l'ensemble de ses utilisateurs.

Des modèles vulnérables à l'adversarial learning

L'apprentissage adversarial *(adversarial machine learning)* est un domaine à l'intersection de la sécurité informatique et de l'intelligence artificielle. Il étudie les vulnérabilités des modèles d'apprentissage automatique face à des attaques malveillantes. Dans les faits, un acteur de la menace peut manipuler des données d'entrée afin de contourner la classification ou de mettre en évidence les limites de décision du modèle LLM attaqué[42]. Cette pratique soulève des préoccupations de sécurité, ces techniques pouvant être utilisées pour attaquer des systèmes d'apprentissage automatiques, et ce même si l'acteur de la menace n'a pas accès au code source.
Le 27 mai 2025, l'entreprise américaine Meta possédant notamment Facebook et Instagram a décidé que toutes les publications, photos, commentaires et interactions publiques des utilisateurs adultes de Facebook et Instagram en Europe seraient collectées pour entraîner ses modèles d'intelligence artificielle[43]. Ce choix expose l'entreprise à un risque accru de compromission de ses modèles par des attaques d'empoisonnement collaboratif, où des acteurs malveillants pourraient volontairement injecter, à grande échelle, des contenus biaisés, trompeurs ou malveillants dans les données publiques collectées.

L'empoisonnement de modèles collaboratifs consiste à manipuler les données d'entraînement de LLMs afin d'induire des prédictions biaisées ou inexactes au modèle. L'adversaire cherche à orienter sélectivement la donnée de sortie du modèle, l'objectif étant d'obtenir des prédictions erronées pour certaines entrées tout en maintenant la précision pour d'autres[44]. Plusieurs schémas d'attaques sont envisagés, dont voici deux exemples :

- **Label Contamination Attacks :** une étiquette (label) d'entraînement désigne l'annotation ou la valeur cible associée à chaque exemple lors de l'apprentissage supervisé.[45] Concernant des tâches de classification, l'étiquette correspond à la catégorie attribuée à un texte, par exemple positif, neutre ou négatif. L'enjeu sera de manipuler ces étiquettes, par exemple en associant une étiquette « négatif » à un avis manifestement positif, afin que le modèle apprenne de mauvaises associations entre les textes et leurs catégories.
- **Decision Time Attack :** cette attaque consiste à modifier de façon ciblée les caractéristiques des données d'entrée pour tromper un modèle d'IA lors de la prédiction du résultat[46]. Les effets concrets de ces attaques sont très variés : elles peuvent provoquer des erreurs de reconnaissance simples, par exemple lorsqu'un modèle confond un panda avec un gibbon, mais aussi permettre des actions plus dangereuses, comme manipuler le comportement de voitures autonomes ou échapper à des systèmes de sécurité basés sur la vidéo, l'audio ou les empreintes digitales[47]. À titre d'exemple, des chercheurs ont piégé l'Autopilot d'une Tesla Model X en projetant brièvement un panneau stop sur un écran publicitaire, provoquant l'arrêt du véhicule[48].



Une image de panda à laquelle on ajoute un bruit imperceptible, faisant croire au modèle qu'il s'agit d'un gibbon. Source : Cybernews[47]

Les méthodes d'attaques adversariales sont applicables aux LLMs collaboratifs comme ChatGPT, notamment lors des phases de fine-tuning ou via des mécanismes de retours d'expérience de la part des utilisateurs. Toutefois, l'efficacité et la faisabilité de ces attaques dépendent fortement des politiques de gestion des données, des contrôles de sécurité, et de la vigilance des opérateurs de ces plateformes. Les attaques adversariales sont théoriquement plus efficaces contre des développements de modèles open-source, la structure du modèle et l'accès aux données d'entraînement étant nécessaires pour parvenir à créer de réels impacts. Si la phase de fine-tuning a pour objectif de rendre les modèles open-source moins exposés aux attaques, les systèmes développés dans un environnement cloisonné restent moins susceptibles d'être compromis.

Conclusion

L'intégration de l'intelligence artificielle dans les modes opératoires malveillants est devenue un élément central d'une part toujours croissante des stratégies d'attaque actuellement observées. Pour autant, leur potentiel demeure restreint par des barrières techniques, bien que celles-ci reculent avec l'amélioration continue des modèles.

L'IA permet néanmoins d'industrialiser des tâches qui nécessitaient une expertise humaine. Les phases de reconnaissance, de collecte d'informations et de cartographie des cibles sont désormais accélérées par des outils bénéficiant des apports de l'IA, capables de traiter et de structurer des volumes massifs de données, offrant ainsi aux acteurs malveillants une représentation globale des vecteurs de compromission exploitables de leurs cibles. Les techniques d'ingénierie sociale ont également bénéficié des progrès rapides de l'IA. Les campagnes de phishing ciblées et automatisées sont désormais modulables à volonté, adaptant leurs discours en fonction des données collectées lors de la phase de reconnaissance. Cette sophistication s'étend aux attaques multicanales, où phishing, vishing et deepfakes vocaux sont orchestrés de manière cohérente, renforçant la crédibilité des scénarios d'ingénierie sociale.

La conception de logiciels malveillants sophistiqués impose de concilier en permanence des exigences de furtivité, de persistance et performance. Or, ces ajustements stratégiques exigent une forme de raisonnement tactique et une flexibilité que les LLM ne possèdent pas à ce jour. Il en va de même pour la capacité de la GenAI à détecter et exploiter des failles de manière autonome[49]. Parallèlement, la multiplication des vulnérabilités ciblant les modèles d'intelligence artificielle et leurs outils de développement expose la recherche et les organisations à de nouveaux risques d'attaques sophistiquées, incluant le vol de données sensibles et la compromission de chaînes d'approvisionnement logicielles.

Les capacités actuelles de l'intelligence artificielle ne suffisent pas à instaurer un écosystème cybercriminel pleinement automatisé, capable de mener sans supervision chacune des phases de la kill chain. Si la tendance se dirige vers une plus grande automatisation, l'intervention humaine demeure requise à la fois pour entraîner, purifier et guider les modèles, mais aussi pour orienter l'IA à chaque étape de la kill chain, depuis la reconnaissance initiale jusqu'au déploiement et à l'adaptation des attaques en fonction des réactions de la cible.

À l'avenir, l'évolution des techniques d'Intelligence Artificielle pourrait toutefois bouleverser ce panorama.

Les progrès en matière d'auto-apprentissage, de génération automatique de code malveillant et d'adaptation dynamique aux contre-mesures laissent entrevoir la possibilité d'attaques plus autonomes, capables de s'auto-adapter sans intervention humaine directe. L'IA pourrait ainsi prendre en charge des décisions tactiques, optimiser les vecteurs d'attaque, voire gérer la résilience des opérations face à la détection ou à la neutralisation. Cette automatisation progressive accroîtrait la rapidité, la discrétion et la capacité d'innovation des attaquants, rendant les menaces plus difficiles à anticiper et à contrer.

À l'horizon 2025 et au-delà, l'intelligence artificielle devrait donc profondément transformer le paysage de la défense en cybersécurité, en réponse à la sophistication croissante des attaques automatisées. Les dispositifs d'analyse prédictive, fondés sur l'exploitation de vastes corpus de données historiques et le traitement en temps réel des flux d'information, pourraient permettre d'identifier précocement les vulnérabilités et d'anticiper l'émergence de menaces[50]. L'IA permettrait alors d'automatiser non seulement la détection des incidents, mais également les réponses aux attaques, en neutralisant rapidement les tentatives de compromission sans intervention humaine directe. L'une des options potentiellement envisagées par les organisations pourrait être d'intégrer massivement des architectures de type Zero Trust, où chaque accès serait validé en continu[51]. Les outils d'analyse comportementale, alimentés par l'IA, favoriseraient une surveillance proactive des endpoints et des flux de données, tout en isolant efficacement les menaces dès leur apparition.

Toutefois, comme ce papier a pu le démontrer, la généralisation de ces technologies ne s'effectuerait pas sans risques : l'IA pourrait elle-même devenir une cible privilégiée des attaquants, qui tenteraient de manipuler les modèles ou de contaminer les données d'entraînement pour contourner les défenses. À l'avenir, la capacité des organisations à anticiper et à se prémunir contre les attaques adversariales s'imposera comme une priorité absolue, la robustesse des modèles d'intelligence artificielle constituant dès lors un enjeu stratégique central pour la cybersécurité de demain. ■

Bibliographie

[1] M. T. Intelligence, «Staying ahead of threat actors in the age of AI,» 2024. [En ligne] **https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/**

[2] «Qu'est-ce que la cyber-kill chain ?,» [En ligne] **https://www.microsoft.com/fr-fr/security/business/security-101/what-is-cyber-kill-chain**

[3] Arxiv, «A Survey on Offensive AI Within Cybersecurity,» 7 Octobre 2024. [En ligne] **https://arxiv.org/pdf/2410.03566**

[4] CMSWIRE, «LinkedIn Scraped by Bad Bots in Massive Scale Attack,» 2016. [En ligne] **https://www.cmswire.com/information-management/linkedin-scraped-by-bad-bots-in-massive-scale-attack/**

[5] WebAsha, «How Hackers Use AI for Reconnaissance : The Role of Artificial Intelligence in Cybersecurity Threats and Data Gathering,» 2025. [En ligne] **https://www.webasha.com/blog/how-hackers-use-ai-for-reconnaissance-the-role-of-artificial-intelligence-in-cybersecurity-threats-and-data-gathering**

[6] Ronantakizawa, «Github project, nmap.ai,» [En ligne] **https://github.com/ronantakizawa/nmap.ai**

[7] S. University, «Automated Social Media Reconnaissance in Modern Cyber Threats.,» 2024.

[8] S. M. Hazim Hanif, «VulBERTa:Simplified Source Code Pre-Training for Vulnerability Detection,» 2022. [En ligne] **https://arxiv.org/abs/2205.12424**

[9] E. Zurich, «Large Language Model for Vulnerability Detection: Emerging Results and Future Directions.,» 2024. [En ligne] **https://arxiv.org/abs/2401.15468**

[10] ResearchGate, «VulBERTa training pipeline. Steps are taken in order from 1 to 8.,» [En ligne] **https://www.researchgate.net/figure/ulBERTa-training-pipeline-Steps-are-taken-in-order-from-1-to-8_fig1_364069820**

[11] PortSwigger, «Burp AI,» [En ligne] **https://portswigger.net/burp/documentation/desktop/burp-ai**

[12] Netenrich, «FraudGPT: The Villain Avatar of ChatGPT,» 2023. [En ligne] **https://netenrich.com/blog/fraudgpt-the-villain-avatar-of-chatgpt**

[13] Darktrace, «Business Email Compromise (BEC) in the Age of AI,» 2024. [En ligne] **https://www.darktrace.com/blog/business-email-compromise-bec-in-the-age-of-ai**

[14] S. A. University, «Lateral Phishing With Large Language Models: A Large Organization Comparative Study,» 2025. [En ligne] **https://www.researchgate.net/publication/390288076_Lateral_Phishing_with_Large_Language_Models_A_Large_Organization_Comparative_Study**

[15] KnownBe4, «Phishing Threat Trends Report,» 2025. [En ligne] **https://www.knowbe4.com/hubfs/Phishing-Threat-Trends-2025_Report.pdf**

[16] T. Micro, «ICO Scams Leverage 2024 Olympics to Lure Victims, Use AI for Fake Sites,» 2024. [En ligne] **https://www.trendmicro.com/en_us/research/24/f/ico-scams-leverage-2024-olympics-to-lure-victims-use-ai-for-fake.html**

[17] FBI, «Business Email Compromise: The $55 Billion Scam,» 2024. Federal Bureau of Investigation, 2024, Business Email Compromise: The $55 Billion Scam (I-091124-PSA).

[18] SlashNext, « The State of Phishing 2024,» 2025. [En ligne] **https://slashnext.com/the-state-of-phishing-2024/**

[19] Sophos, «A familiar playbook with a twist: 3AM ransomware actors dropped virtual machine with vishing and Quick Assist,» 2025. [En ligne] **https://news.sophos.com/en-us/2025/05/20/a-familiar-playbook-with-a-twist-3am-ransomware-actors-dropped-virtual-machine-with-vishing-and-quick-assist/**

[20] Trustpair, « Deepfake : porte ouverte à la fraude en entreprise ?,» 2024. [En ligne] **https://trustpair.com/fr/blog/deepfake-porte-ouverte-a-la-fraude-en-entreprise/**

[21] C. World, « Finance worker pays out $25 million after video call with deepfake 'chief financial officer',» 2024. [En ligne] **https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk**

[22] Group-IB, «Face Off: Group-IB identifies first iOS trojan stealing facial recognition data,» 2024. [En ligne] **https://www.group-ib.com/blog/goldfactory-ios-trojan/**

[23] Toolify.ai, « Create Realistic Face Swaps with SimSwap,» 2024. [En ligne] **https://www.toolify.ai/ai-news/create-realistic-face-swaps-with-simswap-2749440**

[24] Hacksider, «Deep-Live-Cam,» 2023. [En ligne] **https://github.com/hacksider/Deep-Live-Cam**

[25] Deloitte, «Generative AI and the fight for trust,» 2024. [En ligne] **https://www2.deloitte.com/content/dam/Deloitte/us/Documents/Advisory/us-generative-ai-and-the-fight-for-trust.pdf**

[26] T. Micro, «A Closer Look at ChatGPT's Role in Automated Malware Creation,» 2023. [En ligne] **https://www.trendmicro.com/en_us/research/23/k/a-closer-look-at-chatgpt-s-role-in-automated-malware-creation.html**

[27] D. Harry, «LLM-enabled Developer Experience (as of April 2024),» 2024. [En ligne] **https://www.linkedin.com/pulse/llm-enabled-developer-experience-april-2024-drew-harry-h0k4c**

[28] C. f. E. T. a. Security, « Evaluating Malicious Generative AI Capabilities,» 2024. [En ligne] **https://cetas.turing.ac.uk/publications/evaluating-malicious-generative-ai-capabilities**

[29] HP, « HP Wolf Security Uncovers Evidence of Attackers Using AI to Generate Malware,» 2024. [En ligne] **https://www.hp.com/us-en/newsroom/press-releases/2024/ai-generate-malware.html**

[30] Coolaj86, «Chat GPT «DAN» (and other «Jailbreaks»),» 2023. [En ligne] **https://gist.github.com/**