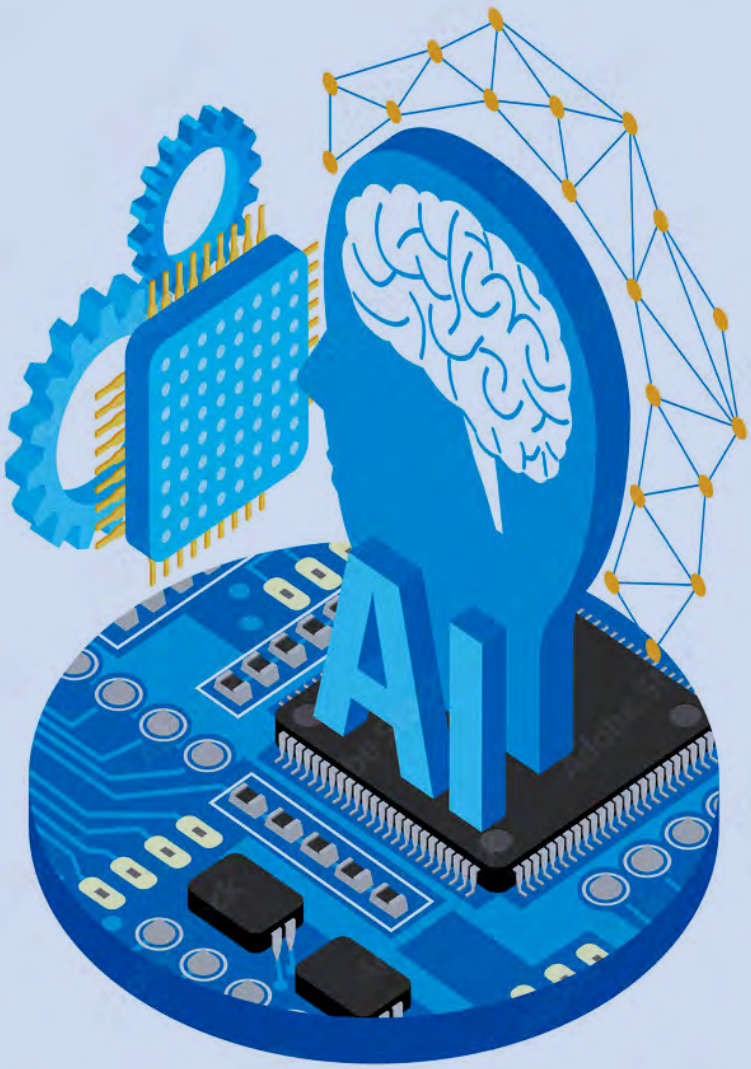
CCN-CERT

BP/30



Approche de l'intelligence artificielle et de la cybersécurité

RAPPORT DE BONNES PRATIQUES

OCTOBRE **2023**

ccn-cert centro criptológico nacional

CCN centro criptológico nacional

Éditeur :

MINISTERIO DE DEFENSA
SUBSECRETARÍA DE DEFENSA
SECRETARÍA GENERAL TÉCNICA

CCN centro criptológico nacional

Paseo de la Castellana 109, 28046 Madrid

Auteurs : Carlos Galán Cordero et Carlos M. Galán Pascual
Date d'émission : octobre 2023

LIMITATION DE LA RESPONSABILITÉ

Le présent document est fourni conformément aux termes qu'il contient, rejetant expressément tout type de garantie implicite qui puisse y être liée. Le Centre National de Cryptologie ne peut en aucun cas être tenu responsable des dommages et préjudices —directs, indirects, fortuits ou extraordinaires— dérivés de l'utilisation de l'information et du logiciel contenus dans ce document, même s'il a été averti de cette possibilité.

AVIS JURIDIQUE

Contenu

Contenu

Objet du document

Comme son titre l'indique, ce document a pour objectif d'examiner de plus près le domaine de travail de deux disciplines : l'intelligence artificielle et la cybersécurité. Malgré leurs origines clairement séparées dans le temps, celles-ci ont été témoins —au cours des dernières années et jusqu'à l'actualité— du rapprochement et de l'unification de leurs domaines d'expertise autour d'une nouvelle activité pratique capable de rassembler les connaissances et l'expérience antérieure des deux : l'**intelligence artificielle au service de la cybersécurité** ou ce que nous pourrions appeler **"l'Intelligence Artificielle pour la Cybersécurité" (*Artificial Intelligence Cybersecurity*, AICS).**

Son caractère introductif, plus proche d'une étude que d'un traité scientifique, lui permet de parvenir plus facilement à ses lecteurs cibles : les professionnels ou les chercheurs des systèmes d'information, de leurs utilisations et de leurs enjeux, et plus particulièrement les dirigeants des organisations (publiques ou privées), leurs départements de gestion —y compris les directions techniques et juridiques— et, bien sûr, les équipes de cybersécurité.

Par ce document, nous espérons fournir un premier aperçu sur ce sujet. Cette étude pourra être complétée par d'autres textes plus spécifiques (telles que ceux référencés ici) ou tout autre document futur qui puisse rapporter des informations sur les nouvelles —et sans doute surprenantes— réalités qui viendront se matérialiser, d'un commun accord, dans le domaine de la cybersécurité et de l'intelligence artificielle.

AVIS DE NON-RESPONSABILITÉ :

Afin de faciliter la compréhension du texte, des appareils, des instruments et du matériel commercial provenant de différentes entités ont dû être identifiés. Cette identification ne signifie pas qu'il existe une recommandation ou une approbation de la part du Centre National de Cryptologie en Espagne (CCN), ni que le matériel ou les appareils identifiés soient nécessairement les meilleurs disponibles pour l'objectif indiqué dans chaque cas.

1. Introduction

1.1 Définition de l'intelligence artificielle (IA) et de la cybersécurité

Bien qu'il n'existe pas de consensus permettant de fournir une définition universelle, nous pouvons dire que l'**intelligence artificielle (IA)** est un sous-domaine de l'informatique qui vise à développer des systèmes capables d'effectuer des tâches qui, jusqu'à présent, requièrent l'intelligence humaine. Ces tâches peuvent inclure l'**apprentissage** (acquisition d'informations et de règles d'utilisation de ces informations), le raisonnement (utilisation de règles pour parvenir à des conclusions approximatives ou définitives) et l'**autocorrection**[1].

Par exemple, l'actuelle proposition de règlement européen sur l'intelligence artificielle, en trilogues[3] au moment de la rédaction du présent document, stipule qu'une définition de l'IA devrait être basée sur les principales caractéristiques fonctionnelles du logiciel et, en particulier, sur sa capacité à générer, par rapport à un ensemble spécifique d'objectifs définis par l'homme, du contenu, des prédictions, des recommandations, des décisions ou d'autres informations de sortie qui influencent l'environnement avec lequel le système interagit, que ce soit

1 L'ENISA, dans son document ARTIFICIAL INTELLIGENCE AND CYBERSECURITY RESEARCH (Jun, 2023) souligne que : "Il n'existe pas de définition commune de l'IA (Commission européenne. Centre commun de recherche. AI watch : defining Artificial Intelligence : towards an operational definition and taxonomy of artificial intelligence. Office des publications, 2020. doi:10.2760/382730 (https://data.europa.eu/doi/10.2760/382730).

Bien qu'il n'existe pas de définition commune, les définitions examinées présentent certains points communs (cf. CCI5) qui peuvent être considérés comme les principales caractéristiques de l'IA : (i) la perception de l'environnement, y compris la prise en compte de la complexité du monde réel ; (ii) le traitement de l'information (collecte et interprétation des données (sous forme de données) ; (iii) la prise de décision (y compris le raisonnement et l'apprentissage) : entreprendre des actions, exécuter des tâches (y compris s'adapter et réagir aux changements dans l'environnement) avec un certain niveau d'autonomie ; (iv) la réalisation d'objectifs spécifiques".

2 Proposition de RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL ÉTABLISSANT DES RÈGLES HARMONISÉES SUR L'INTELLIGENCE ARTIFICIELLE (LOI SUR L'INTELLIGENCE ARTIFICIELLE) ET MODIFIANT CERTAINS ACTES LÉGISLATIFS DE L'UNION (Bruxelles, 21.4.2021).

3 Les trilogues sont des groupes informels constitués pour chaque proposition législative et composés de trois membres : un de la Commission, un du Parlement et un de la présidence du Conseil (https://spanish-presidency.consilium.europa.eu/es/noticias/los-trilogos/).

dans une dimension physique ou numérique. Il faut ajouter qu'une définition de "système d'IA" devrait être complétée par une liste des techniques et stratégies spécifiques utilisées dans son développement.

Le terme **"cybersécurité"** désigne essentiellement la pratique consistant à protéger les systèmes, les réseaux et les programmes contre les cyberattaques. Ces cyberattaques visent souvent à accéder à des informations précieuses ou confidentielles, à les modifier ou à les détruire, à extorquer de l'argent aux utilisateurs ou à perturber les processus et les services.

1.2 Brève histoire de l'IA dans la cybersécurité

La relation entre l'IA et la cybersécurité s'est consolidée au fil des ans. Au départ, les systèmes de cybersécurité s'appuyaient principalement sur des signatures et des règles prédéfinies pour détecter les menaces. Cependant, avec l'augmentation et l'évolution des menaces cyber, le besoin de systèmes plus avancés et adaptatifs est devenu évident.

Les premières tentatives d'utilisation de **techniques d'apprentissage automatique pour la détection des intrusions**[4], ont eu lieu dans les années 1990, mais ce n'est que dans les années 2010, grâce aux progrès des technologies d'**apprentissage profond** et à la disponibilité de grands ensembles de données, que l'IA a commencé à jouer un rôle important dans la cybersécurité, en offrant des solutions plus efficaces et plus précises face à des menaces en constante évolution.

4 En effet, au cours des années 1990, l'utilisation de techniques d'apprentissage automatique pour la détection des intrusions a suscité un intérêt croissant, car il a été reconnu que les techniques traditionnelles basées sur les signatures ne suffiraient pas à détecter les attaques nouvelles ou modifiées, connues sous le nom d'attaques "zero-day". En réponse à cela, des techniques basées sur le comportement et l'apprentissage automatique ont été explorées, telles que : **IDES (Intrusion Detection Expert System)** : développé à la fin des années 1980 et au début des années 1990 par SRI International, IDES a été l'un des premiers systèmes de détection d'intrusion basés sur le comportement. Il utilise des techniques statistiques pour établir un profil de l'activité "normale" d'un utilisateur ou d'un système, puis lance des alertes en cas d'écarts significatifs par rapport à ce comportement. **Système ADAM** : En 1995, un système appelé ADAM (Automated Detection, Analysis, and Measurement) a été proposé par Lee et Stolfo. Ce système utilise des algorithmes de regroupement pour détecter les activités anormales dans les systèmes d'audit. **Réseaux de neurones** : Au cours des années 1990, les réseaux de neurones ont également été étudiés en tant qu'outil de détection des intrusions. Par exemple, en 1998, Ghosh, Schwartzbard et Schatz ont proposé d'utiliser les réseaux de neurones pour détecter les comportements anormaux dans les connexions réseau. **Système LERAD** : à la fin des années 1990, Barbara et Wu ont mis au point le système LERAD (Learning Rules for Anomaly Detection), une technique basée sur l'apprentissage automatique pour détecter les activités anormales dans les ensembles de données d'audit. **Ensemble de données de la KDDD Cup 1999** : La KDDD Cup 1999 est peut-être l'un des efforts les plus influents dans le domaine de la détection des intrusions basée sur l'apprentissage automatique. Elle a fourni un ensemble de données contenant une variété d'intrusions simulées dans un environnement réseau. Cet ensemble de données a été largement utilisé par la communauté de chercheurs pour évaluer et comparer différentes méthodes de détection des intrusions.

PÉRIODES	ACTIVITÉS
Les débuts (ANNÉES 1960 ET 1970) :	Au début du développement de l'informatique, l'idée d'automatiser la sécurité n'était pas une priorité. Les systèmes informatiques n'étaient pas aussi largement interconnectés qu'ils le sont aujourd'hui, et le concept même de cybersécurité n'en était qu'à ses balbutiements. Les premières approches de l'IA au cours de cette période se sont concentrées sur des sujets tels que le traitement du langage naturel et les *systèmes experts*, mais pas sur la cybersécurité.
Naissance de la cybersécurité (ANNÉES 1980) :	Avec l'essor de l'informatique personnelle et le développement des premiers réseaux, les premières menaces cyber sont apparues. Les outils de sécurité s'appuient sur des signatures et des modèles connus pour détecter les menaces, ce qui peut être considéré comme une forme primitive d'apprentissage automatique, bien que l'IA en tant que telle n'ait pas encore été intégrée de manière significative dans la cybersécurité.
Premières approches de l'IA dans le domaine de la cybersécurité (ANNÉES 1990) :	Les systèmes de détection d'intrusion (IDS) ont commencé à intégrer des techniques d'apprentissage automatique de base pour identifier les schémas de trafic anormaux. La recherche et les travaux académiques ont commencé à explorer l'utilisation d'algorithmes de classification pour améliorer la détection des logiciels malveillants et des attaques.
Explosion du Big Data et progrès de l'IA (ANNÉES 2000) :	Avec la prolifération de l'internet et l'émergence de menaces plus avancées, les grands ensembles de données (journaux, trafic réseau, etc.) sont devenus une source cruciale pour la cybersécurité. Des techniques d'intelligence artificielle ont commencé à être utilisées pour analyser ces grands volumes de données afin de détecter des schémas suspects ou des comportements anormaux.
Apprentissage profond et cybersécurité (ANNÉES 2010) :	L'essor de l'apprentissage profond (en particulier les réseaux de neurones convolutifs et récurrents) a trouvé des applications dans la cybersécurité, comme la détection avancée de logiciels malveillants basée sur les caractéristiques et les comportements plutôt que sur les signatures. Des systèmes de réponse automatique et d'orchestration[5], ont été introduits sur le site, utilisant l'IA pour prendre des décisions en temps réel face aux menaces identifiées. Cependant, le concept d'"attaques adverses"[6] est également apparu contre les modèles d'IA. Dans ce type d'attaque, les attaquants cherchent à tromper ou à confondre les modèles d'apprentissage automatique.
Présent et futur (À PARTIR DES ANNÉES 2020) :	Au moment présent, l'IA est un outil essentiel en matière de cybersécurité, non seulement pour la détection et la réponse, mais aussi pour la prédiction des menaces. Les menaces cyber devenant de plus en plus sophistiquées, le besoin de solutions d'IA plus avancées et plus robustes, y compris l'IA générative dans la cybersécurité, se fait également sentir. Les préoccupations relatives à l'éthique, à la vie privée et à la responsabilité de l'IA au service de la cybersécurité sont également mises en avant, et ces domaines sont susceptibles de connaître un développement important dans les années à venir, comme le montre le règlement européen sur l'IA mentionné ci-dessus.

5 *Security Orchestration, Automation, and Response (SOAR) systems.*

6 *Adversarial attacks.*

1.3 Importance actuelle du sujet

Aujourd'hui, la cybersécurité n'est pas seulement une question technique, mais une préoccupation mondiale qui touche les institutions publiques, les entreprises et les particuliers. Avec la numérisation de nombreux services et la création d'infrastructures critiques connectées, la nécessité de sécuriser ces systèmes est primordiale ; tout cela en raison de la réalité posée par les caractéristiques suivantes de la **transformation numérique** de la société :

- **La croissance exponentielle des données :** on peut dire que nous vivons à l'ère du *Big Data*. Des pétaoctets de données sont générés chaque jour et, dans ce vaste océan d'informations, la détection de schémas malveillants ou de comportements anormaux s'avère une tâche extrêmement compliquée. L'IA, grâce à des algorithmes avancés, est capable d'analyser de grands volumes de données en temps réel et d'identifier des menaces potentielles qu'il serait pratiquement impossible de détecter par des méthodes manuelles ou traditionnelles.
- **La constante évolution des menaces :** les menaces cyber ne sont pas statiques, mais en constante évolution. Les acteurs de la menace développent constamment de nouvelles techniques et tactiques pour contourner les systèmes de sécurité. L'IA permet l'adaptabilité et l'apprentissage continu, ce qui signifie qu'elle peut "apprendre" des nouvelles menaces et s'adapter en conséquence, offrant ainsi une couche de protection supplémentaire.
- **L'automatisation et la réponse rapide :** face à une cyberattaque, la réponse doit être immédiate. L'IA peut automatiser les actions à entreprendre, telles que l'isolement d'un appareil compromis ou le blocage d'un accès suspect, beaucoup plus rapidement qu'un humain ne pourrait le faire. Cela réduit le temps d'exposition et minimise potentiellement les dommages.
- **La reconnaissance de modèles complexes :** l'IA est exceptionnellement utile pour identifier des modèles dans de grands ensembles de données. Dans le contexte de la cybersécurité, cela signifie qu'elle peut identifier des comportements malveillants sur la base de modèles subtils qui pourraient passer inaperçus dans les systèmes traditionnels.

Aujourd'hui, la cybersécurité n'est pas seulement une question technique, mais une préoccupation mondiale qui touche les institutions publiques, les entreprises et les particuliers

- **Le manque de professionnels de la cybersécurité :** la demande de professionnels de la cybersécurité ne cesse de croître. L'IA peut contribuer à combler cette lacune, en prenant en charge les tâches qui nécessitent une analyse et une réponse en temps réel et en permettant aux experts humains de se concentrer sur des tâches plus stratégiques.
- **Le coût économique et social :** les failles de sécurité peuvent entraîner d'énormes pertes économiques, nuire à la réputation et, dans le cas des infrastructures critiques, mettre en danger des vies humaines. L'IA appliquée à la cybersécurité ne protège pas seulement les actifs et les données d'une organisation, mais peut également jouer un rôle crucial dans la protection de la société dans son ensemble.
- **Les défis éthiques et réglementaires :** à mesure que l'IA s'intègre dans la cybersécurité (et dans la vie en général), de nouvelles questions éthiques et réglementaires se posent. Qui est responsable si l'IA prend une mauvaise décision ? Comment s'assurer que l'IA agit de manière équitable et non discriminatoire ? Ces questions cruciales soulignent l'importance de considérer l'IA non seulement d'un point de vue technique, mais aussi d'un point de vue éthique et social[7].

L'IA appliquée à la cybersécurité ne protège pas seulement les actifs et les données d'une organisation, mais peut également jouer un rôle crucial dans la protection de la société dans son ensemble

Tout cela sans oublier que, comme le souligne l'ENISA[8], de multiples **acteurs et acteurs de la menace** utilisent déjà des techniques d'IA pour développer leurs actions, parmi eux :

- **Les cybercriminels,** dont la motivation première est le profit, auront tendance à utiliser l'IA comme outil d'attaque, mais aussi à exploiter les vulnérabilités des systèmes d'IA existants. Par exemple, ils pourraient essayer d'attaquer des chatbots d'IA pour voler des informations sur des cartes de crédit ou d'autres données. Ils pourraient également lancer une attaque par ransomware contre des systèmes basés sur l'IA utilisés pour la gestion de la chaîne d'approvisionnement et le stockage.

7 Pour plus d'informations à ce sujet, voir "*La certification comme mécanisme de contrôle de l'intelligence artificielle en Europe*" (C. Galán. Institut espagnol d'études stratégiques. 2019). https://www.ieee.es/Galerias/fichero/docs_opinion/2019/DIEEEO46_2019CARGAL-InteligenciaArtificial.pdf

8 ENISA- AI Cybersecurity Challenges (2021).

- **Les personnes ayant accès à des informations privilégiées**, y compris les employés et les sous-traitants qui ont accès aux réseaux d'une organisation, peuvent effectuer des actions malveillantes, qu'elles soient intentionnelles ou non. Des intrus malveillants peuvent, par exemple, tenter de soustraire ou de saboter l'ensemble de données d'entraînement utilisé par les systèmes d'IA de l'entreprise. En revanche, d'autres personnes peuvent corrompre involontaire ou accidentellement l'ensemble de données d'entraînement.
- **Les États-nations** ou **les acteurs parrainés par un État** qui, en plus de développer des moyens d'exploiter les systèmes d'IA pour attaquer d'autres pays (y compris leurs industries et infrastructures critiques) et d'utiliser des systèmes d'IA pour défendre leurs propres réseaux, rechercheront activement des vulnérabilités dans les systèmes d'IA qu'ils pourront exploiter. Il peut s'agir d'un moyen de nuire à un autre pays ou de recueillir des informations.
- **Les terroristes**, qui cherchent à causer des dommages physiques ou même des pertes en vies humaines, par exemple en cyberattaquant les véhicules sans conducteur pour les utiliser comme une arme.
- **Les hacktivistes**, dont la plupart ont des motivations idéologiques, peuvent également essayer de pirater des systèmes d'IA pour démontrer qu'il est possible de le faire. De plus en plus de groupes s'inquiètent des dangers potentiels de l'IA et il n'est pas rare qu'ils piratent un système d'IA pour obtenir de la publicité.
- Il existe également des **acteurs non sophistiqués**, tels que les script kiddies, qui peuvent avoir des motivations criminelles ou idéologiques. Il s'agit généralement de personnes non qualifiées qui utilisent des scripts ou des programmes pré-écrits pour attaquer les systèmes, car elles n'ont pas les connaissances nécessaires pour les coder elles-mêmes.
- Outre ces acteurs traditionnels de la menace, il semble de plus en plus nécessaire d'inclure les **concurrents** parmi les acteurs de la menace, car certaines entreprises pourraient commencer à manifester clairement leur intention d'attaquer leurs rivaux afin de gagner des parts de marché.

Le paysage des menaces est donc vaste et extrêmement sensible[9].

Des intrus malveillants peuvent, par exemple, tenter de soustraire ou de saboter l'ensemble de données d'entraînement utilisé par les systèmes d'IA de l'entreprise

9 ENISA (2021), *op. cit*.

TAXONOMIE DES MENACES LIÉES À L'IA

L'activité néfaste/abus (*Nefarious activity/abuse, NAA*) : "actions intentionnelles dirigées contre les systèmes, infrastructures et réseaux TIC par le biais d'actes nuisibles dans le but de soustraire, d'altérer ou de détruire une cible spécifique".

L'écoute, l'interception et le détournement (*Eavesdropping/Interception/Hijacking, EIH*) : "actions visant à écouter, interrompre ou prendre le contrôle de la communication d'un tiers sans son consentement".

Les attaques physiques (*Physical Attacks, AP*) : "actions visant à détruire, exposer, altérer, désactiver, soustraire ou obtenir un accès non autorisé à des biens physiques tels que les infrastructures, le matériel (hardware) ou les interconnexions".

Les dommages non intentionnels (*Unintentional Damages, UD*) : actions non intentionnelles qui causent "des destructions, des dommages ou des lésions aux personnes ou aux biens et qui entraînent une défaillance ou une réduction de l'utilité".

Les défaillances/dysfonctionnements (*Failures/Malfunctions, FM*) : "dysfonctionnements partiels ou totaux d'un bien (matériel ou logiciel)".

Les coupures (*Outages, OUT*) : "interruptions inattendues du service ou diminution de la qualité en dessous d'un niveau requis".

Les catastrophes (*Disasters, DIS*) : "accident soudain ou catastrophe naturelle causant des dommages importants ou des pertes de vies humaines".

Juridique *(Legal, LEG)* : "actions en justice menées par des tiers (contractants ou non contractants), dans le but d'interdire des actions ou de compenser des pertes sur la base du droit applicable".

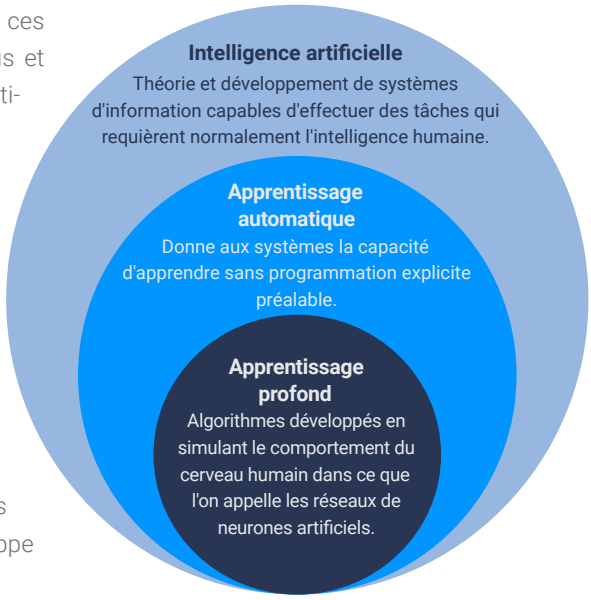
2. Les fondements de l'intelligence artificielle

L'intelligence artificielle (IA) constitue un vaste domaine d'étude qui englobe diverses techniques et technologies. Depuis les débuts de l'informatique jusqu'à nos jours, l'IA est passée d'un concept théorique à un outil pratique qui peut être utilisé dans d'innombrables domaines, dont la cybersécurité. Dans ce contexte de cybersécurité, l'IA agit comme un **multiplicateur de force**, offrant des capacités avancées qui vont au-delà de ce qui est possible avec les méthodes traditionnelles.

Pour comprendre les avantages de l'IA pour la cybersécurité, il est essentiel de se familiariser avec les techniques et technologies spécifiques qui s'utilisent actuellement. Ces techniques vont de l'apprentissage automatique et de ses sous-domaines à la logique floue, aux réseaux de neurones ou à l'IA générative plus récente. Chacune de ces techniques possède ses propres caractéristiques, avantages, défis et applications en matière de cybersécurité et, ensemble, elles constituent un arsenal que les organisations peuvent utiliser pour se défendre contre les croissantes et changeantes menaces cyber.

Dans cette section, nous allons explorer quelques-unes des principales techniques et technologies d'IA utilisées dans le domaine de la cybersécurité, en fournissant une base solide pour comprendre comment l'IA est en train de révolutionner la façon dont nous protégeons nos systèmes et nos données.

D'un point de vue descriptif, il est utile de placer les différents modèles d'IA dans un contexte qui permet de mieux comprendre leurs techniques et leurs caractéristiques. La figure ci-dessous développe cette idée.



2.1 L'apprentissage automatique (*Machine Learning, ML*)

L'apprentissage automatique est une méthode d'analyse des données qui automatise la construction de modèles analytiques. Au lieu d'être explicitement programmées pour effectuer une tâche, les machines sont "entraînées" en utilisant de grands ensembles de données et en exécutant des algorithmes qui leur donnent la capacité d'*apprendre* comment effectuer la tâche.

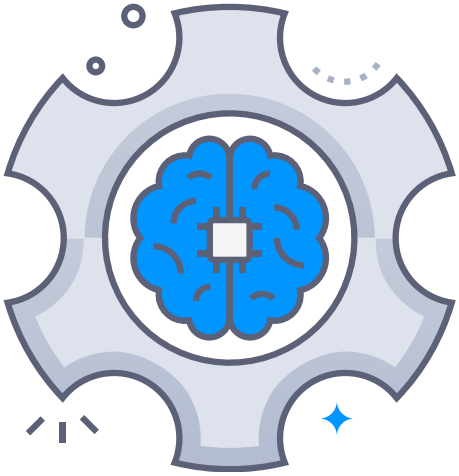
Les techniques d'apprentissage automatique peuvent être classées selon les **types** suivants :

Apprentissage supervisé	Il s'agit de la technique la plus courante. Le modèle est entraîné à l'aide d'un ensemble de données étiquetées, ce qui signifie que chaque exemple de l'ensemble de données est accompagné de la "bonne réponse". Une fois entraîné, le modèle peut commencer à faire des prédictions ou à prendre des décisions sans intervention humaine. Parmi les exemples d'applications, on peut citer la classification des courriers électroniques en "spam" ou "non-spam" ou la prédiction des prix des logements sur la base de caractéristiques telles que la taille et l'emplacement.
Apprentissage non supervisé	Dans ce cas, le modèle est formé sur un ensemble de données non étiquetées et son objectif est de découvrir des structures cachées dans les données. Les techniques courantes comprennent le regroupement et la réduction de la dimensionnalité. Un exemple pourrait être la segmentation des clients en différents groupes, sur la base de leur comportement d'achat.
Apprentissage par renforcement	Il s'agit d'un type d'apprentissage dans lequel un agent apprend à se comporter dans un environnement donné en effectuant certaines actions et en recevant des récompenses ou des pénalités en réponse. Il est souvent utilisé dans les domaines de la robotique, des jeux et de la navigation.

Au sein de ces types, des techniques spécifiques ont été développées, telles que les **arbres de décision** (*decision trees*), les **machines à vecteurs de support** (*support vector machines*), le **classificateur de Bayes** (*Naive Bayes'classifier*), le **regroupement** dit **K-means**, le **modèle de Markov caché** ou les **algorithmes génétiques**, que nous examinerons brièvement ci-dessous.

Dans le domaine de la cybersécurité, l'apprentissage automatique peut être utile pour la **détection des menaces** (car il permet d'analyser de grands volumes de données pour identifier des modèles de comportements anormaux ou suspects, ce qui permet une détection plus rapide et plus précise des menaces), l'**analyse de code** (puisqu'en entraînant des modèles d'IA sur des ensembles de données de logiciels malveillants, l'apprentissage automatique peut aider à identifier et à classer de nouvelles variantes, même si elles n'ont pas été observées auparavant) et la **détection du phishing et de la fraude** (en demandant à des modèles d'apprentissage automatique d'analyser les caractéristiques des sites web et des courriels, afin de déterminer s'ils sont malveillants ou légitimes).

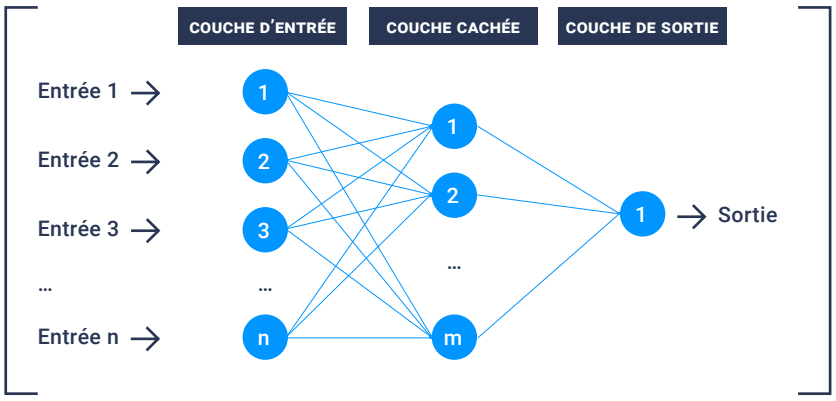
Toutefois, la mise en place de systèmes de cybersécurité fondés sur l'intelligence artificielle comporte des exigences et des défis, tels que la **disponibilité de données de bonne qualité**, car l'apprentissage automatique ne vaut que les données sur lesquelles il est entraîné, de sorte que si les données sont biaisées ou de mauvaise qualité, les modèles qui en résultent seront également médiocres ; ou ce que l'on appelle le **surajustement** (*overfitting*), car un modèle peut être trop complexe et "mémoriser" l'ensemble des données d'entraînement, au lieu de généraliser sur des données nouvelles et inédites ; ou ce que l'on appelle les **attaques adverses**, par lesquelles un attaquant peut essayer de tromper un modèle d'apprentissage automatique en présentant des données spécifiquement conçues pour le dérouter et générer des décisions erronées.



2.2 L'apprentissage profond (*Deep Learning, DL*)

L'apprentissage profond est un procédé d'apprentissage automatique utilisant des **réseaux de neurones**[10] possédant plusieurs couches (trois ou plus).

Les réseaux de neurones artificiels (ANN) sont des modèles informatiques inspirés du fonctionnement des neurones du cerveau humain. Ils sont composés d'unités ou de nœuds appelés "neurones" qui sont organisés en couches : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque connexion entre les neurones possède un poids associé, qui est affiné au cours du processus d'entraînement.



Ces réseaux sont capables d'apprendre des modèles et des représentations de données à des niveaux de complexité croissants, ce qui leur permet d'effectuer des tâches qui étaient considérées comme trop complexes pour les algorithmes traditionnels d'apprentissage automatique.

Nous pouvons classer ces technologies selon les **types** suivants :

10 Plusieurs ensembles de données d'entraînement sont actuellement disponibles. Les plus utilisés sont : KDD Cup99 ; DEFCON ; CTU-13 ; Spam Base ; SMS Spam Collection ; CICIDS2017 ; CICAndMal2017 ; Android Validation ; IoT-23 data set ; chacun ayant des caractéristiques particulières pour répondre à des problèmes spécifiques.

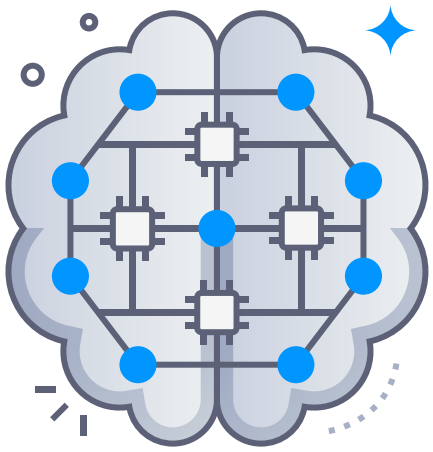
Réseaux de neurones convolutifs (CNN)	Ils sont particulièrement utiles pour la reconnaissance d'images, la reconnaissance de formes et/ou la vision par ordinateur, car ils permettent d'identifier efficacement les motifs dans une image.
Réseaux de neurones récurrents (RNN)	Ils sont particulièrement efficaces lorsqu'on utilise des séquences de données, telles que des séries chronologiques ou des textes, en raison de leur capacité à "mémoriser" les informations précédentes de la séquence.
Réseaux de neurones à mémoire à long terme (LSTM)	Il s'agit d'une variante des RNN, conçue pour résoudre le problème de l'évanouissement du gradient[11] et conserver les informations à long terme. Comme les RNN, ils sont utilisés pour l'analyse de séquences, mais avec une plus grande précision sur les séquences plus longues.
Réseaux adversaires génératifs (GAN)	Il s'agit d'un modèle qui utilise deux réseaux (un réseau génératif et un réseau discriminatif) qui travaillent ensemble pour générer des données d'apparence authentique.

Dans le domaine de la cybersécurité, ces techniques de DL peuvent être utilisées pour :

- **Détection de codes malveillants :** vu que les réseaux de neurones peuvent être entraînés à identifier les logiciels malveillants sur la base de modèles et de caractéristiques extraits de fichiers. Par exemple, un réseau CNN pourrait analyser le contenu d'un fichier binaire et déterminer s'il présente ou non des caractéristiques de logiciel malveillant (malware).
- **Analyse du trafic réseau :** les RNN, compte tenu de leur capacité à analyser des séquences, peuvent être utiles pour inspecter le trafic réseau à la recherche de schémas anormaux ou malveillants.
- **Détection du phishing :** les CNN peuvent être entraînés à analyser le contenu visuel des sites web et à déterminer s'ils imitent ou reproduisent des sites légitimes, dans le but de tromper les utilisateurs.

11 Le problème de l'évanouissement du gradient (*vanishing gradient*) est un obstacle qui survient lors de l'entraînement des réseaux de neurones artificiels traditionnels, en particulier les réseaux de neurones récurrents (RNN). Il s'agit de la tendance des gradients à diminuer de manière exponentielle au fur et à mesure de leur rétropropagation entre les couches et au fil du temps dans les RNN. Lorsque les gradients approchent de zéro, cela signifie que les poids des neurones ne sont pas mis à jour de manière efficace au cours du processus d'entraînement, ce qui conduit à un entraînement inefficace ou au point mort. Les réseaux à mémoire à long terme (LSTM) ont été spécialement conçus pour résoudre ce problème, ainsi que le problème connexe de l'explosion du gradient, où les gradients peuvent croître de manière exponentielle. Les LSTM y parviennent grâce à leur structure cellulaire, qui comprend une porte d'entrée, une porte d'oubli et une porte de sortie. Ces portes, combinées à une cellule d'état, permettent aux LSTM de conserver ou d'éliminer des informations sur de longues séquences, ce qui garantit que le gradient soit préservé et qu'il se propage correctement dans le réseau sans s'estomper ni exploser. Cette conception spéciale permet aux LSTM d'apprendre les dépendances à long terme dans les données, ce qui les rend particulièrement utiles pour des tâches telles que la traduction automatique, le traitement du langage naturel, la prédiction de séries temporelles et autres, où il est crucial de se souvenir d'informations provenant de parties antérieures d'une séquence.

- **Génération d'échantillons de logiciels malveillants pour les essais :** les GAN peuvent être utilisés pour créer des échantillons de logiciels malveillants. Ainsi, la composante générative crée des échantillons tandis que la composante discriminante évalue leur authenticité, ce qui peut contribuer à améliorer la robustesse de certains outils.
- **Analyse comportementale :** les réseaux de neurones peuvent apprendre des modèles de comportement de l'utilisateur ou du système et détecter les écarts par rapport à ces modèles, ce qui pourrait indiquer une activité malveillante ou une compromission.



Toutefois, comme dans le cas précédent, l'utilisation de ces techniques nécessite la prise en compte de certaines questions, telles que la **nécessité de disposer de grands ensembles de données** (étant donné que l'apprentissage profond nécessite souvent de grandes quantités de données étiquetées), le **temps d'entraînement** (étant donné que l'apprentissage des modèles d'apprentissage profond peut être intensif en termes de calcul) ou l'**interprétabilité** (étant donné que, contrairement à d'autres algorithmes, les réseaux de neurones profonds agissent souvent comme des "boîtes noires", ce qui signifie que leurs décisions ne sont pas facilement interprétables par l'homme).

2.3 Algorithmes de classification

Les algorithmes de classification sont une branche de l'apprentissage automatique supervisé. Leur principal objectif est d'attribuer une classe prédéfinie à une entrée (ou exemple) donnée. Dans le contexte de la cybersécurité, ces classes pourraient être, par exemple, "malveillant" ou "bienveillant".

Ainsi, un algorithme de classification vise à apprendre, à partir d'un ensemble de données d'entraînement, à classer les entrées non mappées dans une ou plusieurs catégories (classes).

Nous pouvons classer les algorithmes de tri selon les **types** suivants, en notant leurs applications dans le domaine de la cybersécurité.

TYPE	CARACTÉRISTIQUES	CAS D'USAGES EN CYBERSÉCURITÉ
Régression logistique	Méthode statistique permettant d'analyser des ensembles de données dans lesquels une ou plusieurs variables indépendantes déterminent un résultat. Le résultat est mesuré à l'aide d'une variable dichotomique (oui/non, 1/0, vrai/faux).	**Déterminer si une activité est malveillante ou non en fonction de plusieurs caractéristiques.**
Machines à vecteur de support (SVM)	Ces algorithmes cherchent à trouver l'hyperplan permettant d'obtenir la plus nette séparation possible d'un ensemble de données en classes.	**Classification des courriels en tant que spam ou non-spam, détection des logiciels malveillants basée sur des caractéristiques.**
Arbres de décision et forêts aléatoires	Les arbres de décision divisent l'ensemble des données en sous-ensembles en fonction de la valeur des caractéristiques d'entrée. Les forêts aléatoires sont une collection d'arbres de décision qui participent ensemble à la production d'une prédiction finale.	**Détection des intrusions sur la base de caractéristiques telles que : le type de protocole, l'adresse IP, la durée, etc.**
Réseaux de neurones	Ils ont déjà été étudiés auparavant. Il s'agit de structures inspirées du cerveau humain, capables d'apprendre des schémas complexes.	**Détection de logiciels malveillants, analyse du comportement des utilisateurs, détection d'anomalies, entre autres.**
Algorithme des k plus proches voisins (KNN ou k-NN)	Il s'agit d'un algorithme qui classe une entrée en fonction du classement de ses k plus proches voisins.	**Détection d'activités malveillantes en partant de l'hypothèse que des points similaires peuvent être trouvés les uns à côté des autres (ici, des comportements connus).**
Naive Bayes ou Classification naïve bayésienne	Algorithme basé sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Il est particulièrement utile lorsque la dimension des données est élevée.	**Classification des courriels en spam ou légitimes, analyse de texte pour identifier les communications malveillantes.**

Comme toujours, l'application pratique de ces techniques nécessite la prise en compte de certaines questions, telles que la **qualité des données** (car, comme nous l'avons déjà souligné, la qualité d'un modèle dépend de celle des données sur lesquelles il est entraîné), le **déséquilibre des classes** (car, dans de nombreux scénarios de cybersécurité, tels que la détection de logiciels malveillants, la majorité des exemples peuvent être bienveillants et seul un petit pourcentage malveillant, ce qui peut entraîner un biais de modèle s'il n'est pas traité correctement), l'**interprétabilité** (car il est important de comprendre les raisons qui sous-tendent les décisions d'un modèle, en particulier dans un contexte critique tel que la cybersécurité), ou l'**adaptabilité** (car si les acteurs de la menace évoluent constamment dans leurs tactiques, techniques et procédures, il est essentiel que les modèles de classification puissent s'adapter rapidement à de nouvelles menaces).

2.4 L'IA générative

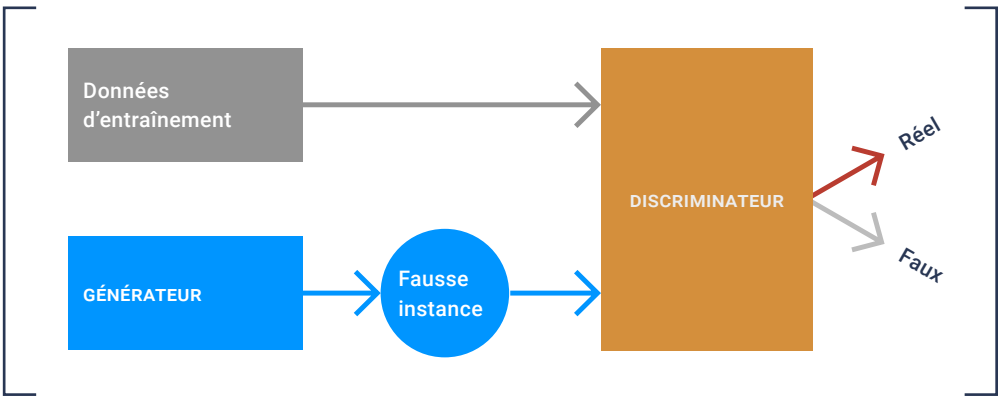
L'intelligence artificielle générative (IA) fait référence à un sous-ensemble de techniques d'apprentissage automatique qui visent à générer de nouvelles données similaires, mais non identiques, aux données sur lesquelles elles ont été entraînées. Contrairement aux techniques d'apprentissage automatique discriminatoires, qui apprennent à différencier parmi les types de données (par exemple, classer les courriels comme "spam" ou "non-spam"), les techniques génératives visent à produire des données qui ressemblent aux données d'entrée.

L'une des approches les plus populaires et les plus efficaces de l'IA générative est celle des **réseaux adversaires génératifs (GAN)**, que nous avons déjà mentionnée. Ces réseaux sont basés sur deux modèles qui fonctionnent ensemble :

1. Un élément **générateur**, qui a pour but la création de données. Au départ, il produit des données de manière aléatoire, mais avec le temps et la rétroaction du discriminateur, il améliore sa capacité à générer des données qui ressemblent à des données réelles.
2. Un élément **discriminateur**, qui examine les données et tente de faire la distinction entre les données réelles et les données générées par l'élément générateur. Il fournit un retour d'information au générateur sur son efficacité (ou son manque d'efficacité).

Ainsi, l'élément générateur tente de produire de fausses données de plus en plus convaincantes, tandis que l'élément discriminateur améliore constamment sa capacité à détecter ces fausses données. Avec un entraînement suffisant, le générateur peut finalement produire des données qui sont presque impossibles à différencier des données réelles pour les humains et les machines.

L'intelligence artificielle générative (IA) fait référence à un sous-ensemble de techniques d'apprentissage automatique qui visent à générer de nouvelles données similaires, mais non identiques, aux données sur lesquelles elles ont été entraînées

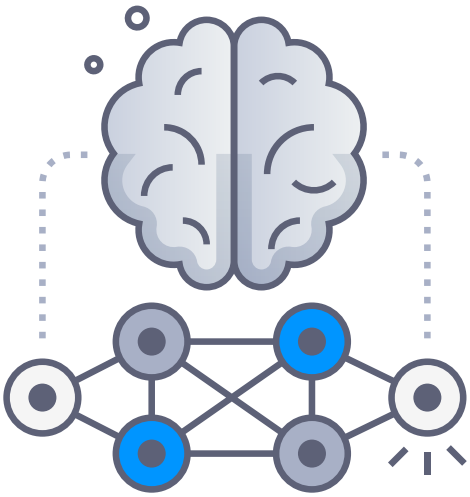


Les applications de l'IA générative sont multiples : **Traitement d'images** (création d'images artistiques, amélioration de la résolution d'images, génération d'images d'objets ou de scènes qui n'existent pas dans la réalité, etc.) ; **Audio** (création de musique, d'effets sonores ou de voix synthétiques) ; **Texte** (génération de textes cohérents, d'histoires, de poèmes, etc.) ; **Vidéo** (création de vidéos synthétiques ou modification de vidéos existantes, telles que les deepfakes, etc.) ; **Données synthétiques** (génération d'ensembles de données pour l'entraînement lorsque les données réelles sont inexistantes ou insuffisantes) ; **Conception et modélisation** (génération de conceptions pour les produits, l'architecture ou la modélisation 3D, etc.).

Voici quelques applications de l'IA générative en matière de cybersécurité :

- **Création d'échantillons de logiciels malveillants :** les GAN peuvent être entraînés à générer des variantes de logiciels malveillants qui échappent à la détection par les solutions de sécurité traditionnelles. Bien que cela puisse sembler dangereux, les chercheurs en sécurité peuvent utiliser cette technique dans des environnements contrôlés pour améliorer la robustesse des systèmes de détection.
- **Renforcement des systèmes de détection :** en générant des logiciels malveillants ou du trafic réseau malveillant, les équipes de sécurité peuvent utiliser ces échantillons pour entraîner et améliorer leurs systèmes de détection. En fait, l'IA est dressée contre elle-même pour améliorer la détection.

- **Simulation du trafic réseau :** l'IA générative peut simuler le trafic réseau normal ou le trafic d'attaque pour tester la robustesse d'un réseau ou d'un système. Cela est particulièrement utile pour former les cyberdéfenseurs et tester les systèmes de sécurité.
- **Création de faux domaines :** dans le domaine de la protection contre les menaces de type "zero-day", les GAN peuvent être utilisés pour générer de faux domaines qui ressemblent à de vrais domaines malveillants. Cela permet aux systèmes de sécurité de prévoir et de bloquer les domaines qui pourraient être utilisés dans de futures attaques.
- **Attaques adverses :** comme mentionné précédemment, les attaques adverses impliquent l'introduction de petites perturbations dans les données afin de tromper les modèles d'apprentissage automatique. Les GAN peuvent être utilisés pour générer efficacement ces perturbations, ce qui peut aider les défenseurs à comprendre et à atténuer ces attaques.
- **Phishing et génération de contenu malveillant :** les GAN peuvent être entraînés à générer des courriels ou des pages web qui imitent des courriels légitimes, ce qui en fait des outils potentiellement utiles pour les attaques de phishing. Toutefois, ils peuvent également être utilisés à des fins défensives, en générant des échantillons de phishing pour entraîner les systèmes de détection.



3. Cas d'usages de l'IA en cybersécurité

Comme chacun sait, la cybersécurité est une lutte sans fin entre les attaquants et les défenseurs. Tandis que les attaquants cherchent de nouvelles vulnérabilités et des moyens de compromettre les systèmes, les défenseurs cherchent à anticiper (prévention), à détecter (détection) et à répondre à ces attaques (réponse).

L'IA, avec sa capacité à traiter de grandes quantités de données à une vitesse extraordinaire et à en tirer des enseignements, offre des solutions significatives aux défis (actuels et émergents) de la cybersécurité.

Le tableau ci-dessous présente quelques-unes des principales applications de l'IA dans le domaine de la cybersécurité :

Détection des menaces et réaction	Les systèmes basés sur l'IA peuvent analyser les schémas du trafic réseau ou le comportement des utilisateurs afin d'identifier les anomalies ou les activités suspectes. Une fois celles-ci détectées, l'IA peut agir rapidement, souvent plus vite qu'une équipe humaine, pour atténuer ou neutraliser la menace.
Analyse prédictive	L'IA peut utiliser des données historiques pour prédire les menaces ou les vulnérabilités futures, ce qui permet aux organisations de se préparer et de se protéger de manière proactive[12].
Authentification et gestion de l'identité	L'IA peut utiliser des données biométriques avancées, le comportement de l'utilisateur et d'autres facteurs pour authentifier les individus avec une grande précision, réduisant ainsi le risque d'accès non autorisé.
Protection contre le phishing	En analysant le contenu, les images et les modèles de textes ou de documents (par exemple les courriels), l'IA peut identifier les tentatives de phishing avec une grande précision, protégeant ainsi les utilisateurs contre les escroqueries potentielles.
Optimisation des paramètres de sécurité	L'IA peut évaluer les configurations et les politiques de sécurité afin d'identifier les faiblesses éventuelles et de proposer des améliorations.

12 De plus amples informations sur l'analyse prédictive et le modèle diamant peuvent être trouvées dans le Guide CCN-STIC 425 Le cycle du renseignement et l'analyse d'intrusions. https://www.ccn-cert.cni.es/series-ccn-stic/guias-de-acceso-publico-ccn-stic/1093-ccn-stic-425-ciclo-de-inteligencia-y-analisis-de-intrusiones/file.html

Dans cette section, nous examinerons en détail la place de l'IA dans ces domaines et dans d'autres domaines de la cybersécurité, son potentiel et, bien sûr, les considérations éthiques et de protection de la vie privée liées à son utilisation.

3.1 Détection des menaces et analyse comportementale

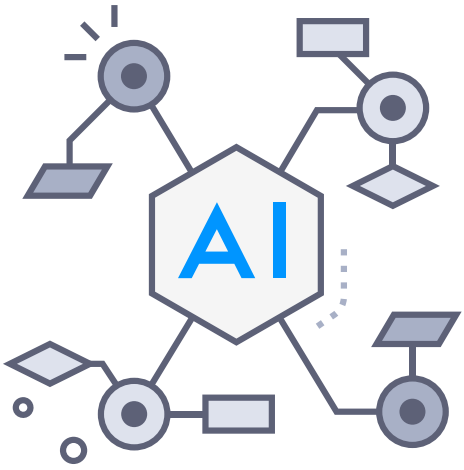
La détection des menaces et l'analyse comportementale sont essentielles pour identifier les cyberattaques et y répondre en temps réel. Avec la mise en œuvre de l'IA dans ces domaines, la cybersécurité a connu une nette amélioration de l'efficacité et de la précision de la détection.

La quantité de données que les organisations (publiques ou privées) traitent quotidiennement est immense. Il est pratiquement impossible de détecter manuellement les menaces dans de tels volumes. De leur côté, les cyberattaques modernes utilisent souvent des tactiques furtives, telles que le déplacement latéral et la persistance, ce qui les rend difficiles à détecter avec les méthodes traditionnelles.

Ainsi, au lieu de s'appuyer uniquement sur les signatures connues des logiciels malveillants, l'IA se concentre sur les **modèles de comportement anormal**. Cela permet de détecter des menaces inconnues jusqu'alors ou des variantes de logiciels malveillants qui ont été légèrement modifiées. En analysant le comportement de l'utilisateur et du système, l'IA peut identifier des activités inhabituelles, telles que l'accès à des fichiers à des heures indues ou le transfert inhabituel de grandes quantités de données.

La détection comportementale des menaces a connu une croissance rapide en termes de popularité et d'adoption, et un certain nombre d'outils et de systèmes, aussi bien commerciaux qu'open source, spécialisés dans cette approche ont vu le jour. Certains des outils les plus populaires sont énumérés ci-dessous :

- **Darktrace**[13] : Darktrace utilise des algorithmes d'apprentissage automatique et d'IA pour détecter, répondre et atténuer les menaces en temps réel sur la base de modèles de comportement anormal. L'outil est connu pour son "*Enterprise Immune System*", qui apprend et établit ce qui peut être considéré comme une "situation normale" dans le réseau et identifie ensuite les écarts par rapport à cette norme.
- **Vectra**[14] : Vectra offre une détection des menaces en temps réel grâce à l'IA. Il se concentre sur la détection des comportements malveillants dans le trafic réseau et fournit une vue détaillée de la chaîne d'attaque en cours, permettant aux équipes de sécurité de réagir rapidement.
- **CrowdStrike Falcon**[15] : CrowdStrike est connu pour ses solutions de protection des points d'extrémité. Sa plateforme Falcon utilise des techniques élaborées à partir de l'approche comportementale pour détecter et prévenir les menaces que d'autres systèmes basés sur les signatures pourraient manquer.
- **Cylance**[16] : CylancePROTECT est une solution de protection des points d'extrémité qui utilise des modèles d'IA pour identifier et bloquer les logiciels malveillants en fonction de leurs caractéristiques et de leurs comportements, plutôt que des signatures connues.
- **Gurucul**[17] : Fournit des solutions d'analyse comportementale des utilisateurs et des entités (UEBA) en utilisant des algorithmes d'apprentissage automatique pour détecter les menaces internes, les fraudes et les accès non autorisés.
- **Wazuh**[18] : Il s'agit d'une plateforme open source pour la détection des menaces, la gestion des vulnérabilités et la surveillance de l'intégrité. Elle utilise des règles et des décodeurs pour analyser les événements de sécurité et détecter les comportements anormaux.
- **Snort**[19] : Bien qu'il soit plus connu en tant que système de détection et de prévention des intrusions (SDPI), Snort a évolué pour incorporer des capacités basées sur le comportement. La communauté Snort développe et partage de nouvelles règles capables de détecter des comportements anormaux.



13 https://es.darktrace.com/
14 www.vectra.ai
15 www.crowdstrike.com
16 www.cylance.com
17 www.gurucul.com
18 www.wazuh.com
19 www.snort.org

- **ELK Stack** (Elasticsearch, Logstash, Kibana)[20] : Bien qu'ELK ne soit pas en soi un outil de détection comportementale, il peut être configuré avec des plug-ins et des règles spécifiques pour effectuer une analyse comportementale des logs et des événements.

Par ailleurs, les systèmes d'IA fonctionnant selon le modèle de l'**apprentissage automatique pour l'analyse comportementale** sont entraînés à l'aide de vastes ensembles de données de comportements, qu'elles soient légitimes ou malveillantes. Grâce à l'apprentissage supervisé, l'IA peut apprendre à classer et à détecter les activités anormales. Ainsi, au fil du temps et au fur et à mesure que les données sont traitées, ces systèmes peuvent améliorer leur précision à l'aide de l'apprentissage non supervisé et l'apprentissage par renforcement.

Les systèmes d'IA fonctionnant selon le modèle de l'apprentissage automatique pour l'analyse comportementale sont entraînés à l'aide de vastes ensembles de données de comportements, qu'elles soient légitimes ou malveillantes

De nombreux outils de cybersécurité modernes ont intégré l'apprentissage automatique (ML) dans leurs capacités afin d'améliorer la détection et la réponse aux menaces. Ces outils utilisent l'apprentissage automatique pour apprendre et s'adapter aux nouvelles menaces en étudiant les modèles et les comportements dans les données. Outre les outils énumérés ci-dessus, certaines des solutions les plus populaires sont présentées ci-dessous :

- **Endgame**[21] : Cette plateforme utilise le ML pour la protection des points d'extrémité, la détection des menaces et la réponse. Sa capacité de ML se concentre sur la détection des techniques et tactiques d'attaque sans s'appuyer uniquement sur les signatures.
- **PatternEx**[22] : Il s'agit d'une solution d'analyse du comportement des utilisateurs et des entités (UEBA) qui utilise l'apprentissage automatique. Elle analyse de grands volumes de données afin d'identifier des modèles qui suggèrent une activité malveillante.
- **SentinelOne**[23] : Il s'agit d'une solution de protection des points d'extrémité qui utilise l'apprentissage automatique pour détecter, classer et réagir face aux comportements malveillants et anormaux.

20 www.elastic.co

21 (Acquis par Elastic) : https://www.elastic.co

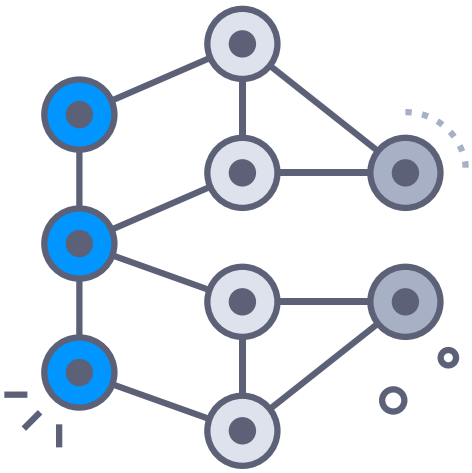
22 https://www.patternex.com

23 https://www.sentinelone.com

- **Kaspersky Machine Learning for Anomaly Detection (MLAD)**[24] : Conçu pour les systèmes industriels, le MLAD de Kaspersky utilise l'apprentissage automatique pour détecter les déviations dans le fonctionnement des machines industrielles.
- **Splunk**[25] : Bien que Splunk soit principalement un outil d'analyse de données et de SIEM (gestion des informations et des événements de sécurité), il possède des capacités qui permettent aux utilisateurs de mettre en œuvre des modèles d'apprentissage automatique afin d'identifier des modèles et des anomalies dans de grands volumes de données.

Les réseaux de neurones, en particulier les réseaux de neurones d'apprentissage profond, se sont révélés efficaces pour détecter des modèles dans de grands ensembles de données. Ils peuvent être utilisés pour identifier les logiciels malveillants dans les fichiers en fonction de leurs caractéristiques, détecter les attaques DDoS en fonction des modèles de trafic ou identifier les tentatives de phishing grâce à l'analyse du texte et du contenu.

Outre les entreprises commerciales mentionnées ci-dessus, qui travaillent également sur les réseaux de neurones, certains des outils les plus connus utilisant ces techniques sont énumérés ci-dessous :



- **Deep Instinct**[26] : Cette société utilise des réseaux de neurones d'apprentissage profond pour prévenir et détecter les logiciels malveillants en temps réel, et propose des solutions pour les points d'extrémité et les appareils mobiles.
- **SparkCognition**[27] : Cette société propose DeepArmor, une solution qui utilise des réseaux de neurones pour fournir une protection contre les menaces en temps réel.
- **NVIDIA**[28] : Bien qu'il ne s'agisse pas d'un outil de sécurité en soi, NVIDIA propose des plateformes et des bibliothèques —telles que CUDA et cuDNN— qui accélèrent les calculs effectués dans les réseaux de neurones.

24 https://www.kaspersky.com
25 https://www.splunk.com
26 https://www.deepinstinct.com
27 https://www.sparkcognition.com
28 https://www.nvidia.com

D'autre part, il existe de nombreux outils qui ont su intégrer des mécanismes d'IA dans leurs technologies, basées sur des concepts plus traditionnels, afin de les rendre plus efficaces.

Voici quelques-unes de ces **applications pratiques traditionnelles :**

Systèmes de détection et de prévention des intrusions (SPDI)	Grâce à l'IA, ces systèmes peuvent détecter et bloquer le trafic malveillant en temps réel avec une plus grande précision. Un système de détection et de prévention des intrusions (SPDI) est essentiel pour détecter les activités malveillantes sur un réseau ou un système et y répondre. L'intégration de l'intelligence artificielle (IA) dans ces systèmes a considérablement amélioré leur capacité à identifier les menaces et à y réagir en temps réel. En voici quelques exemples : • **Darktrace** (https://www.darktrace.com) **:** Comme indiqué ci-dessus, Darktrace est connue pour son approche de la détection et de la prévention des menaces basée sur l'IA. Sa technologie *Enterprise Immune System* utilise l'apprentissage automatique pour détecter les comportements anormaux en temps réel. • **Vectra** (https://www.vectra.ai) **:** Vectra Cognito utilise l'IA pour détecter et hiérarchiser automatiquement les comportements anormaux en temps réel afin de découvrir les attaques actives et les menaces internes. • **Cisco Stealthwatch** (https://www.cisco.com) **:** Bien qu'il ne s'agisse pas d'un SPDI au sens traditionnel, Stealthwatch utilise l'apprentissage automatique pour détecter les comportements anormaux dans le réseau et s'intègre à d'autres solutions Cisco pour fournir des capacités de prévention. • **Lastline** (https://www.lastline.com) **:** Offre des solutions qui utilisent des techniques d'IA, telles que l'apprentissage automatique, pour détecter et répondre aux menaces avancées, évasives et de type "zero-day". • **Awake Security** (https://www.awakesecurity.com) **:** Sa plateforme utilise l'IA pour analyser le trafic réseau et détecter les menaces. Elle peut identifier les comportements malveillants et à risque sans s'appuyer sur des signatures ou des connaissances préalables. • **Fortinet** (https://www.fortinet.com) **:** Fortinet propose toute une série de solutions de sécurité, mais son FortiGate, avec sa fonctionnalité SDPI intégrée, a également intégré l'IA pour améliorer la détection des menaces.

Outils d'analyse criminelle

L'IA peut accélérer les enquêtes à la suite d'un incident de sécurité en identifiant rapidement les indicateurs de compromission et en mappant le parcours d'un attaquant.

La criminalistique numérique, en particulier lorsqu'elle est appliquée aux incidents de sécurité (voir la discipline DEFIR : *Digital Forensics and Incident Response*), peut générer de grandes quantités de données à étudier. L'intelligence artificielle et, en particulier, l'apprentissage automatique (ML) jouent un rôle important dans ce domaine, en aidant à identifier des modèles, à effectuer des analyses plus rapides et à obtenir des informations plus précises.

Voici quelques-uns des outils les plus populaires qui intègrent l'IA en matière d'analyse forensique :

- **Autopsy** (https://www.sleuthkit.org/autopsy/) : Bien qu'il s'agisse avant tout d'un outil d'analyse forensique numérique, il dispose de modules et de plugins qui peuvent exploiter des capacités basées sur l'IA pour analyser les données et rechercher des modèles spécifiques.
- **Cellebrite** (https://www.cellebrite.com) : Connue pour ses solutions d'analyse forensique des appareils mobiles, Cellebrite utilise l'IA pour faciliter l'identification et la catégorisation des données pertinentes sur les appareils mobiles.
- **Brainspace** (https://www.brainspace.com) : il s'agit d'une plateforme d'analyse et de visualisation qui utilise l'apprentissage automatique pour faciliter les enquêtes, l'examen des documents et l'analyse des données. Elle est utilisée dans les enquêtes juridiques, mais peut également être appliquée à la criminalistique numérique.
- **Cyber Triage** (https://www.cybertriage.com) : Associé à Autopsy, cet outil utilise des techniques d'intelligence artificielle pour évaluer rapidement les systèmes compromis, à la recherche de preuves d'activités malveillantes.
- **Endgame** (https://www.elastic.co) : Sa plateforme fournit des capacités de réponse aux incidents et aux menaces et utilise des techniques de ML pour analyser les données et détecter les activités malveillantes.
- **ReversingLabs** (https://www.reversinglabs.com) : Fournit des solutions pour l'analyse des fichiers et artefacts malveillants avec des capacités basées sur l'IA pour identifier, classer et désagréger les menaces.

Ces outils, combinés à l'expertise humaine, peuvent fournir une analyse forensique plus rapide et plus précise, ce qui est crucial pour la réponse aux incidents et les enquêtes.

Systèmes de réponse automatisés	Lorsqu'une menace est détectée, l'IA peut lancer des actions prédéfinies pour contenir ou atténuer l'attaque, par exemple en isolant un système compromis ou en bloquant une adresse IP suspecte. La réponse automatisée, souvent associée à la détection des menaces, est un élément crucial de la sécurité moderne. Grâce à l'intelligence artificielle (IA), ces systèmes peuvent prendre des décisions en temps réel pour contenir, atténuer ou neutraliser les menaces sans intervention humaine immédiate. Outre les entreprises commerciales déjà mentionnées, les outils et solutions les plus populaires qui intègrent l'IA pour fournir des capacités de réponse automatisée sont énumérés ci-dessous : • **Darktrace Antigena** (https://www.darktrace.com) : Antigena est une extension du système de détection de Darktrace basé sur l'IA, qui a la capacité de prendre des mesures automatiques en réponse aux menaces détectées, telles que le blocage des connexions ou la mise en quarantaine des appareils. • **Palo Alto Networks - Cortex XDR** (https://www.paloaltonetworks.com) : Cette plateforme détecte les menaces et automatise la réponse. Elle utilise des techniques d'apprentissage automatique pour identifier les menaces et elle peut, par exemple, bloquer des processus malveillants ou mettre à jour les règles de pare-feu de façon automatique. • **FireEye Helix** (https://www.fireeye.com) : il s'agit d'une plateforme de sécurité qui utilise l'IA pour détecter les menaces et automatiser les réponses. Elle peut s'intégrer à une variété d'outils et de systèmes pour exécuter des actions de réponse. • **IBM Resilient** (https://www.ibm.com) : Il s'agit d'une plateforme de réponse aux incidents qui, associée à *Watson*, le système d'IA d'IBM, peut fournir des recommandations et automatiser des actions en réponse à des incidents de sécurité. • **Fortinet FortiResponder** (https://www.fortinet.com) : Il s'agit d'une solution de réponse aux incidents qui s'intègre à d'autres produits Fortinet pour fournir des capacités automatisées et basées sur des règles. Alors que la réponse est principalement basée sur des règles définies, la détection et les informations peuvent être renforcées par des techniques d'intelligence artificielle. Comme il semble logique de le supposer, l'automatisation doit être utilisée avec précaution. Une mauvaise configuration ou un manque de supervision appropriée peut entraîner des réponses indésirables qui ont un impact négatif sur les opérations. L'IA et l'automatisation doivent être considérées comme des outils qui complètent, mais ne remplacent pas, les experts en sécurité humains.

L'utilisation de ces techniques pose également des défis. Si l'IA peut améliorer la précision, il existe toujours un risque de **faux positifs** qui, s'ils sont nombreux, peuvent entraîner la fatigue de l'équipe de sécurité et d'éventuelles omissions.

Comme pour les systèmes basés sur les signatures, les attaquants développent des techniques pour **échapper à la détection basée sur l'IA**, telles que l'empoisonnement des données ou la manipulation des modèles.

En effet, l'évasion des systèmes basés sur l'IA est une tactique employée par les acteurs de la menace pour éviter d'être détectés par les systèmes de sécurité utilisant des techniques d'intelligence artificielle ou d'apprentissage automatique. Ces méthodes reposent sur la compréhension et l'exploitation des faiblesses ou des biais inhérents aux modèles d'apprentissage automatique. Certains outils et techniques sont conçus spécifiquement à cette fin, tandis que d'autres ont été adaptés pour échapper à l'IA.

Voici quelques-uns des concepts, outils et ressources qui ont été utilisés ou étudiés en rapport avec l'**évasion de l'IA** :

APPRENTISSAGE AUTOMATIQUE ADVERSARIAL	Il s'agit d'une catégorie plutôt que d'un outil spécifique. Les attaques adverses visent à introduire de petites perturbations dans les données d'entrée afin de tromper les modèles d'apprentissage automatique.
CLEVERHANS[29]	Il s'agit d'une bibliothèque logicielle qui fournit des outils pour tester la robustesse des modèles d'apprentissage automatique contre les attaques adverses.
DEEP-PWNING[30]	Il s'agit d'un outil d'évaluation de la sécurité destiné à faciliter l'analyse des systèmes utilisant l'apprentissage profond. Il peut être utilisé pour évaluer la résilience des modèles contre les modifications adverses.
GAN (*GENERATIVE ADVERSARIAL NETWORKS* OU RÉSEAUX ANTAGONISTES GÉNÉRATIFS)	Bien qu'il ne s'agisse pas d'outils d'évasion à proprement parler, les GAN peuvent être utilisés pour générer des données permettant de tromper les systèmes d'IA. Comme indiqué ci-dessus, ces réseaux sont composés de deux éléments : un générateur qui crée des images et un discriminateur qui tente de faire la distinction entre les images réelles et les images générées.
FGSM (*FAST GRADIENT SIGN METHOD*)	Il s'agit d'une technique d'attaque adverse qui introduit des perturbations dans les données d'entrée afin d'embrouiller le modèle d'apprentissage automatique.

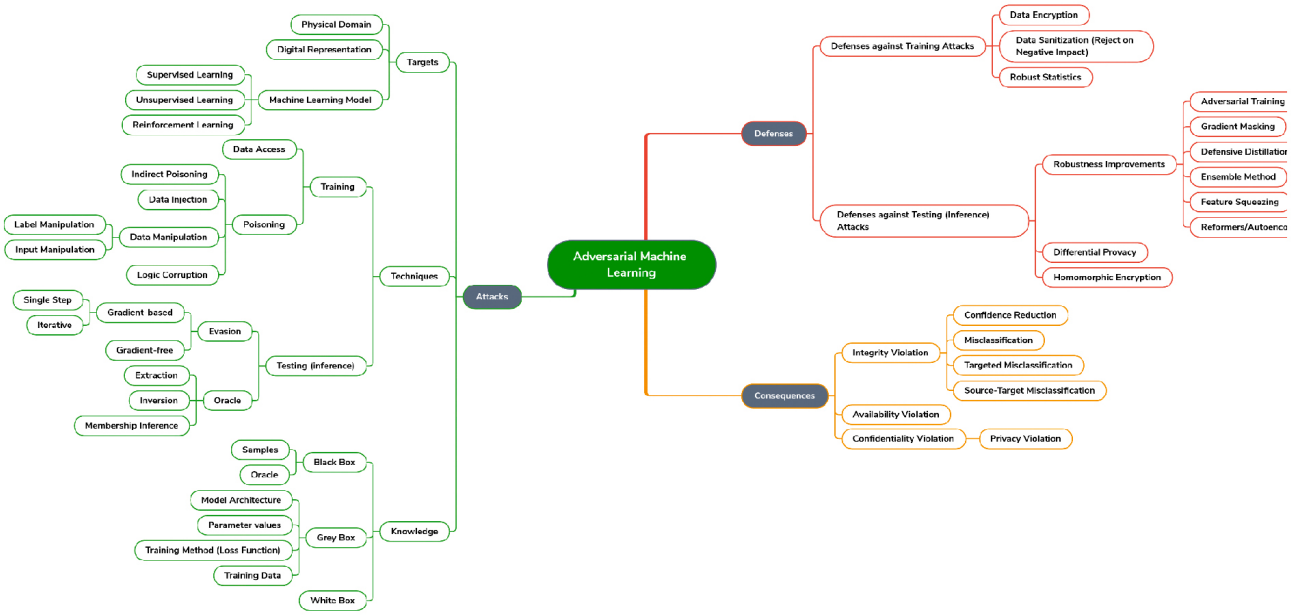
Ces défis de sécurité comprennent, comme nous l'avons vu, l'énorme potentiel de la manipulation adverse des données d'entraînement et de l'exploitation adverse des sensibilités du modèle pour perturber les performances de classification et de régression de ML.

29 https://github.com/tensorflow/cleverhans

30 https://github.com/cchio/deep-pwning

Ainsi, l'AML (*Adversarial Machine Learning*) fait référence à la conception d'algorithmes de ML qui peuvent résister aux défis de sécurité, à l'étude des capacités des attaquants et à la compréhension des conséquences des attaques. Étant donné que les attaques sont lancées par des adversaires mal intentionnés, la sécurité de l'apprentissage automatique doit prendre en compte les défenses visant à prévenir ou à atténuer les conséquences de ces attaques. Bien que les composants de ML puissent également être affectés par divers facteurs involontaires, tels que les défauts de conception ou les biais de données, ces facteurs ne constituent pas des attaques adverses intentionnelles et n'entrent pas dans le champ d'application de la sécurité abordé par la littérature sur l'AML.

Pour son intérêt certain, nous reproduisons ci-dessous la **Taxonomie des attaques, des défenses et des conséquences en matière d'AML**, tirée du National Institute Of Standards and Technology (NIST)[31].



Enfin, il semble évident que l'adoption de l'IA pour la détection des menaces et l'analyse comportementale est en passe de transformer la capacité des organisations à se défendre contre les cyberattaques. Toutefois, comme pour tout outil, il est essentiel de l'utiliser en conjonction avec d'autres techniques et approches de cybersécurité pour assurer une défense complète.

31 NIST - Draft NISTIR 8269 - A Taxonomy and Terminology of Adversarial Machine Learning (2019).

3.2 Réponse automatique et orchestration

L'orchestration, l'automatisation et la réponse de sécurité (SOAR) désignent la capacité d'un système de sécurité à détecter et à répondre automatiquement à une menace ou à une vulnérabilité sans intervention humaine, en coordonnant souvent plusieurs systèmes et outils au cours du processus.

Ce modèle comporte donc trois composantes : **l'orchestration**, c'est-à-dire la coordination et la gestion intégrée des outils et des systèmes de sécurité, leur permettant de travailler ensemble de manière harmonieus[32] ; **l'automatisation**, c'est-à-dire la capacité d'effectuer des tâches spécifiques sans intervention humaine[33] ; et **la réaction**, c'est-à-dire les mesures prises en réponse à un événement de sécurité, qui peuvent être automatiques (par exemple, le blocage d'une adresse IP) ou nécessiter une intervention humaine (par exemple, l'investigation en cas de suspicion d'intrusion).

L'utilisation d'outils configurés selon le modèle SOAR est particulièrement utile car, compte tenu de la vitesse de propagation des menaces, la capacité à réagir automatiquement aux menaces peut être cruciale pour minimiser les dommages.

En outre, l'automatisation permet aux équipes de sécurité de se concentrer sur des tâches de plus grande valeur ou sur des menaces plus sophistiquées, en laissant les tâches répétitives ou routinières aux solutions automatisées, en éliminant le facteur humain de la gestion de ces dernières et, par conséquent, en réduisant le risque d'erreurs ou d'incohérences.

L'orchestration, l'automatisation et la réponse de sécurité (SOAR) désignent la capacité d'un système de sécurité à détecter et à répondre automatiquement à une menace ou à une vulnérabilité sans intervention humaine, en coordonnant souvent plusieurs systèmes et outils au cours du processus

32 Par exemple, si un système détecte un code malveillant (malware), l'orchestration pourrait garantir que l'information est partagée avec tous les outils pertinents en vue d'une analyse et d'une réponse ultérieures.

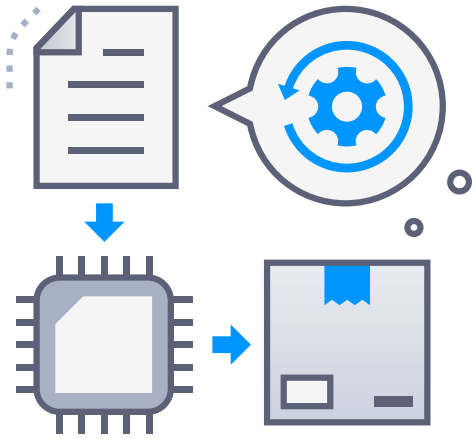
33 Il peut s'agir de bloquer une adresse IP malveillante, de désactiver des comptes d'utilisateurs compromis ou même de corriger des logiciels vulnérables.

Enfin, les systèmes SOAR peuvent traiter un grand nombre d'alertes et d'événements, dont beaucoup seraient insurmontables pour une équipe humaine.

Malgré leurs avantages, les outils SOAR posent également des défis : par exemple, une réponse automatique mal configurée peut causer plus de problèmes qu'elle n'en résout, notamment en bloquant le trafic légitime ou en ignorant les menaces réelles ; ou encore la dépendance excessive à l'égard de ces systèmes sans contrôle ou surveillance humaine, qui peut entraîner une absence de détection des menaces plus sophistiquées ou plus complexes.

Les outils SOAR sont généralement utilisés dans les **scénarios** suivants :

- **Réponse aux incidents :** si un système détecte un comportement anormal, tel qu'une augmentation inhabituelle du trafic vers une destination spécifique, il peut automatiquement bloquer cette activité et alerter l'équipe de sécurité.
- **Intégration des outils :** en intégrant plusieurs outils (tels que les systèmes de détection d'intrusion, les pare-feu et les solutions de sécurité des points d'extrémité), l'orchestration permet une réponse plus holistique et coordonnée face aux menaces.
- **Automatisation des flux de travail :** Par exemple, en cas de détection d'un logiciel vulnérable, un système SOAR pourrait automatiquement lancer un processus de correction ou de mise à jour.



Certains des outils SOAR les plus populaires sont énumérés ci-dessous :

- **Splunk Phantom** (https://www.splunk.com/en_us/software/splunk-security-orchestration.html)
- **Siemplify** (https://www.siemplify.co/)
- **Palo Alto Networks - Cortex XSOAR** (https://www.paloaltonetworks.com/cortex/xsoar)
- **IBM Resilient** (https://www.ibm.com/security/incident-response/resilient-soar-platform)
- **CyberSponse** (https://www.cybersponse.com/)

3.3 Prédiction sur les menaces

Le concept "prédiction sur les menaces" est une évolution de la détection traditionnelle des menaces et représente un changement dans l'approche de la cybersécurité, qui passe d'une position réactive à une position proactive. Par conséquent, la prédiction des menaces fait référence au processus d'anticipation et de détection des cyberattaques potentielles ou des vulnérabilités avant qu'elles ne se matérialisent, en utilisant des techniques d'analyse avancées et, pour la partie qui nous intéresse maintenant, des techniques d'intelligence artificielle pour identifier des modèles et des signaux qui suggèrent une attaque imminente ou l'émergence de nouvelles vulnérabilités.

Le concept "prédiction sur les menaces" est une évolution de la détection traditionnelle des menaces et représente un changement dans l'approche de la cybersécurité, qui passe d'une position réactive à une position proactive

Les avantages de ce modèle sont clairs : il semble évident qu'en prévoyant une menace avant qu'elle ne se produise, les organisations ont le temps de se préparer, de renforcer leurs défenses et de réduire l'impact potentiel de l'attaque. En outre, la prédiction permet aux organisations de concentrer leurs efforts et leurs ressources sur les menaces les plus susceptibles de se produire, plutôt que de les répartir sur un large éventail de scénarios possibles. Enfin, en anticipant les menaces et en agissant de manière proactive, les organisations peuvent améliorer leur position globale en matière de sécurité.

Les outils de ce modèle utilisent essentiellement les **trois méthodes de prédiction** suivantes :

Analyse des tendances	En analysant les tendances passées, les organisations peuvent anticiper les types de menaces qui pourraient apparaître à l'avenir.
Renseignements sur les menaces	Il s'agit de collecter et d'analyser des informations sur les menaces existantes et émergentes à partir de diverses sources, telles que les flux de renseignements, les rapports d'enquête et les données relatives aux événements de sécurité.
Modèles prédictifs	Ils utilisent des algorithmes et des modèles mathématiques, souvent alimentés par l'IA et l'apprentissage automatique, pour analyser de grands ensembles de données et identifier des modèles qui suggèrent une menace imminente.

Comme pour les modèles précédents, ce concept présente également des défis importants.

En effet, la prédiction des menaces, en particulier lorsqu'elle est basée sur des modèles prédictifs, peut conduire à des **faux positifs**, ce qui peut détourner les ressources et l'attention d'autres domaines critiques. D'autre part, la création et la maintenance de modèles prédictifs, en particulier ceux qui utilisent des techniques avancées d'intelligence artificielle, peuvent être **complexes** et nécessiter du personnel spécialisé. Enfin, comme toujours, la précision des prédictions dépend fortement de la **qualité, de la pertinence et de la mise à jour des données d'entrée.**

La prédiction des menaces, en particulier lorsqu'elle est basée sur des modèles prédictifs, peut conduire à des faux positifs, ce qui peut détourner les ressources et l'attention d'autres domaines critiques

Les applications de ce modèle dans le domaine de la cybersécurité se sont concentrées sur la **prévision des codes malveillants** (sur la base des caractéristiques des logiciels malveillants connus, il est possible de prévoir de nouvelles variantes ou évolutions des logiciels malveillants), la **prévision des attaques de phishing** (en analysant les schémas des campagnes de phishing précédentes, il est possible d'anticiper les attaques futures ou d'identifier les domaines suspects) et la **prévision des attaques DDoS** (en observant les schémas de trafic et d'autres signaux, il est possible d'anticiper une attaque DDoS avant qu'elle ne se produise).

Voici quelques-uns des exemples les plus connus d'outils qui ont utilisé ce modèle, dans ses différentes variantes, de l'analyse statistique traditionnelle à l'apprentissage automatique et à l'intelligence artificielle, pour anticiper les menaces avant qu'elles ne se produisent. Certains d'entre eux ont déjà été mentionnés plus haut.

- **Darktrace Antigena** (https://www.darktrace.com/en/products/antigena/)
- **Recorded Future** (https://www.recordedfuture.com/)
- **Palo Alto Networks – AutoFocus** (https://www.paloaltonetworks.com/cortex/autofocus)
- **CyberInt** (https://www.cyberint.com/)
- **Lookout** (https://www.lookout.com/)
- **SparkCognition DeepArmor** (https://www.sparkcognition.com/deeparmor-endpoint-protection/)
- **CrowdStrike Falcon** (https://www.crowdstrike.com/es-es/products/falcon-platform/)
- **CylancePROTECT** (https://www.blackberry.com/us/en/products/blackberry-protect)
- **Plateforme de sécurité Kenna** (https://www.kennasecurity.com/platform/)

3.4 Identification et authentification biométrique

L'identification et l'authentification biométrique font référence à l'utilisation de caractéristiques physiques ou comportementales uniques d'une personne pour vérifier ou confirmer son identité. Ces caractéristiques peuvent comprendre, entre autres, les empreintes digitales, la reconnaissance faciale, la reconnaissance vocale et les motifs de l'iris.

Nous pouvons différencier les **types de biométrie** suivants :

Biométrie physique	Elle est basée sur des caractéristiques physiques du corps, telles que : • **Empreintes digitales :** les crêtes et les vallées au bout des doigts propres à chaque personne. • **Reconnaissance faciale :** la structure et les caractéristiques du visage. • **Reconnaissance de l'iris :** les motifs uniques de l'iris de l'œil. • **Géométrie de la main :** la forme et la taille de la main.
Biométrie comportementale	Elle est basée sur les actions effectuées par l'individu, telles que : • **Dynamique de la frappe :** La manière dont un individu appuie sur les touches d'un clavier. • **Reconnaissance vocale :** les caractéristiques uniques de la voix d'une personne. • **Schéma de marche :** la façon dont une personne marche.

L'utilisation de méthodes biométriques présente un certain nombre d'avantages, tels que **l'unicité** (les caractéristiques biométriques sont propres à chaque individu, ce qui réduit la probabilité de duplication ou d'usurpation), **la commodité** (les utilisateurs n'ont pas besoin de se souvenir de mots de passe ou de codes PIN) et **la difficulté de falsification** (étant donné qu'il est difficile de reproduire ou de falsifier les données biométriques, en particulier par rapport aux mots de passe).

Toutefois, l'utilisation de mécanismes biométriques pose certains **défis et certaines limites**. Par exemple :

- **Erreurs de reconnaissance :** aucun système biométrique n'est précis à 100 %. Il peut y avoir des faux positifs (reconnaître quelqu'un qui n'est pas l'utilisateur) ou des faux négatifs (ne pas reconnaître l'utilisateur légitime).
- **Craintes liées au respect de la vie privée :** la collecte, le stockage et l'utilisation de données biométriques soulèvent des craintes en matière de respect de la vie privée, de consentement et de conformité légale.
- **Irrévocabilité :** contrairement aux mots de passe, qui peuvent être modifiés, les caractéristiques biométriques sont permanentes. Si les données biométriques sont compromises, elles ne peuvent être ni remplacées ni modifiées.
- **Coût :** la mise en œuvre de systèmes biométriques peut nécessiter du matériel et des logiciels spécialisés, ce qui peut entraîner des coûts supplémentaires.

Néanmoins, l'utilisation de la biométrie a trouvé diverses **applications dans le domaine de la cybersécurité**, telles que **l'accès logique sécurisé** (de nombreux appareils et applications offrent des options d'authentification biométrique comme couche de sécurité supplémentaire), **les transactions en ligne** (l'authentification biométrique peut être utilisée dans les transactions bancaires en ligne et les paiements mobiles pour vérifier l'identité de l'utilisateur) ou **le contrôle d'accès physique** (les systèmes biométriques peuvent être utilisés pour contrôler l'accès à certains bâtiments, pièces ou zones à accès limité, etc.).

Néanmoins, l'utilisation de la biométrie a trouvé diverses applications dans le domaine de la cybersécurité, telles que l'accès logique sécurisé, les transactions en ligne ou le contrôle d'accès physique

Comme indiqué précédemment, l'identification et l'authentification biométriques sont devenues populaires dans de nombreux dispositifs en raison de leur capacité à fournir une couche de sécurité supplémentaire. Certains des exemples les plus connus d'outils et de systèmes utilisant la biométrie sont présentés ci-dessous :

- **Apple Face ID et Touch ID :** Face ID permet de déverrouiller l'iPhone, l'iPad et certains Mac à l'aide de la reconnaissance faciale, tandis que Touch ID utilise l'empreinte digitale (Face ID et Touch ID).
- **Windows Hello :** fonctionnalité de Windows 10 qui permet aux utilisateurs d'accéder à leurs appareils par reconnaissance faciale ou d'empreintes digitales. (https://www.microsoft.com/es-es/windows/windows-hello)

- **Samsung Pass :** est un outil d'authentification biométrique qui permet aux utilisateurs d'appareils Samsung de déverrouiller leurs smartphones et d'accéder à des applications et à des sites web en utilisant la reconnaissance de l'iris, du visage ou des empreintes digitales. (https://www.samsung.com/global/galaxy/apps/samsung-pass/)
- **BioID :** est une plateforme d'authentification faciale basée sur le cloud qui peut être intégrée dans diverses applications pour fournir une authentification biométrique. (https://www.bioid.com/)
- **AuthenTrend :** offre des solutions d'authentification basées sur les empreintes digitales pour différentes applications, des clés USB aux solutions d'entreprise. (https://www.authentrend.com/)
- **Nuance VocalPassword:** solution de reconnaissance vocale qui vérifie l'identité de l'utilisateur en se basant sur les caractéristiques uniques de sa voix. (https://www.nuance.com/omni-channel-customer-engagement/security/vocalpassword.html)

Il ne s'agit là que de quelques exemples classiques des nombreuses solutions d'authentification biométrique disponibles sur le marché. Il est essentiel de garder à l'esprit que lorsqu'une solution biométrique est envisagée, il est décisif d'évaluer la sécurité, la confidentialité et la facilité d'utilisation afin de s'assurer qu'elle répond aux exigences spécifiques de l'organisation, de l'utilisateur ou de la réglementation[34].

En outre, les solutions d'identification et d'authentification biométriques ont commencé à intégrer des **capacités d'IA avancées**, en particulier dans des domaines tels que la reconnaissance faciale et l'analyse comportementale, afin d'améliorer la précision et de réduire les faux positifs. Certains des exemples les plus connus sont énumérés ci-dessous :

- **Trueface :** utilise l'IA pour fournir des solutions de reconnaissance faciale. Ses algorithmes apprennent et s'améliorent au fil du temps, ce qui augmente la précision de l'identification. (https://www.trueface.ai/)
- **Kairos :** il s'agit d'une plateforme basée sur le cloud qui utilise l'IA pour analyser les visages dans les vidéos et les photos, offrant des solutions de reconnaissance faciale. (https://www.kairos.com/)
- **BehavioSec :** cette plateforme utilise l'IA pour analyser les modèles de comportement en temps réel, tels que la dynamique de la frappe et la manipulation de la souris, afin d'authentifier les utilisateurs. (https://www.behaviosec.com/)

34 Ce serait le cas de la conformité avec le Schéma d'accréditation Esquema Nacional de Seguridad (ENS) (Décret Royal 311/2022 du 3 mai) pour les entités relevant de son champ d'application.

- **ID R&D :** utilise l'IA dans ses solutions biométriques vocales et comportementales pour fournir une authentification plus sûre et plus efficace. (https://www.idrnd.net/)
- **Deepware Scanner :** est un scanner d'empreintes digitales basé sur des réseaux de neurones profonds. Il utilise l'IA pour analyser et vérifier les empreintes digitales avec une grande précision.
- **Affectiva :** bien que principalement axée sur l'interprétation des émotions par l'analyse faciale, Affectiva utilise l'IA pour l'analyse en temps réel des expressions faciales, ce qui présente des applications potentielles dans les domaines de l'authentification comportementale ou des réponses émotionnelles. (https://www.affectiva.com/)

3.5 Analyse des vulnérabilités et pentesting automatisé

Comme chacun sait, **l'analyse des vulnérabilités** est un processus systématique d'évaluation, d'identification et de classification des failles de sécurité dans les systèmes d'information. Ces vulnérabilités peuvent être causées par des erreurs logicielles, des configurations inadéquates, des défaillances matérielles ou même de mauvaises pratiques de gestion de la sécurité.

Ce processus d'analyse comprend généralement **l'identification** (des outils analysent les systèmes, les réseaux et les applications à la recherche de vulnérabilités connues), **la classification** (une fois détectées, les vulnérabilités sont classées en fonction de leur gravité et du risque qu'elles présentent), **la remédiation** (des solutions sont proposées pour atténuer ou corriger les vulnérabilités détectées) et **la vérification** (après la remédiation, une nouvelle vérification est effectuée pour confirmer que les vulnérabilités ont été traitées de manière adéquate).

Le test de pénétration, communément appelé ***pentesting,*** est une attaque simulée sur un système dans le but de découvrir les vulnérabilités avant que les vrais attaquants ne le fassent. Contrairement à l'analyse des vulnérabilités, qui utilise généralement des analyses automatisées pour identifier les vulnérabilités connues, le *pentesting* implique souvent des experts qui tentent activement d'exploiter les vulnérabilités et de pénétrer dans les systèmes, en simulant les tactiques, techniques et procédures (TTP) des véritables adversaires.

Le processus comprend généralement **la reconnaissance** (collecte d'informations sur la cible), **le balayage** (identification des points d'entrée possibles), **la pénétration** (exploitation des vulnérabilités), **la maintenance de l'accès** (simulation des mouvements d'un attaquant après avoir obtenu l'accès) et **l'analyse** (contenant le rapport des résultats et des recommandations pour fortifier le système).

Comme on peut s'y attendre, **l'intelligence artificielle** a également été intégrée dans l'analyse des vulnérabilités et les tests de pénétration, avec les procédures suivantes :

Amélioration de l'automatisation	Grâce à l'IA, les outils peuvent analyser les réseaux et les systèmes plus rapidement et avec plus de précision, en identifiant les vulnérabilités que les outils traditionnels pourraient manquer.
Apprentissage continu	Les outils basés sur l'IA peuvent apprendre de chaque analyse, en s'adaptant aux nouvelles vulnérabilités et techniques d'attaque.
Simulation avancée	Dans le cadre du pentesting, l'IA peut simuler le comportement d'attaquants plus complexes, en testant les systèmes contre les menaces émergentes et avancées.
Hiérarchisation des risques	L'IA peut aider à hiérarchiser les vulnérabilités en fonction du contexte et des données historiques, ce qui permet aux équipes de sécurité de se concentrer sur les menaces les plus imminentes ou les plus nuisibles.
Intégration et corrélation	Les solutions basées sur l'IA peuvent corréler des données provenant de sources multiples, offrant ainsi une vision plus holistique de la posture de sécurité d'une organisation.

Des outils tels que **Tenable.io, Qualys Cloud Platform** ou **Checkmarx** utilisent déjà des capacités d'IA pour améliorer leur balayage et leur analyse. En outre, les plateformes de *pentesting* telles que **Cobalt** sont en train d'intégrer l'IA pour automatiser et améliorer certaines parties du processus.

L'intégration de l'IA dans ces domaines est prometteuse, mais il est essentiel de rappeler que, pour l'instant, la combinaison d'experts humains et de ces outils avancés constitue l'approche la plus solide et la plus complète de la cybersécurité.

3.6 Défense contre les adversaires automatisés

Au fur et à mesure que la technologie progresse, les défenseurs améliorent leurs outils, mais aussi les attaquants. **Les adversaires automatisés** sont des programmes, des bots et des scripts conçus pour mener des attaques sans intervention humaine directe. Ces attaques peuvent aller de simples attaques par force brute à des modèles plus sophistiqués capables de s'adapter et de changer de tactique à la volée.

Le tableau suivant présente une **typologie** des adversaires automatisés les plus courants et leurs **caractéristiques** générales.

TYPES		CARACTÉRISTIQUES
Bots et scrapers	Ils peuvent être utilisés pour effectuer de nombreuses tâches, telles que le *scraping* de sites web, mais ils peuvent également être utilisés pour des attaques, telles que les tentatives de connexion ou l'exploitation des vulnérabilités d'un site web.	**1. Vitesse :** ils peuvent lancer des attaques à une vitesse pratiquement impossible pour un humain. **2. Adaptabilité :** certains systèmes automatisés avancés peuvent changer de tactique s'ils détectent qu'une approche particulière ne fonctionne pas. **3. L'échelle :** Ils sont capables de traiter des milliers, voire des millions de cibles simultanément. **4. Persistance :** ils peuvent poursuivre leurs attaques pendant de longues périodes sans fatigue ni distraction.
Vers	Il s'agit de programmes malveillants qui se répliquent automatiquement pour se propager à d'autres ordinateurs, souvent en exploitant des vulnérabilités dans les logiciels.	
Bots de DDoS	Certains réseaux de machines zombies (botnets) sont utilisés pour lancer des attaques DDoS coordonnées. Ils inondent les cibles avec un trafic important dans l'objectif d'interrompre le service ou le fonctionnement d'infrastructures.	
Systèmes automatisés de phishing	Ils peuvent rapidement créer des sites web frauduleux ou envoyer des courriels en masse avec des liens malveillants.	

La défense par l'intelligence artificielle

Pour se protéger contre ces adversaires automatisés, la défense doit également être agile, adaptable et, dans de nombreux cas, automatisée. C'est là que l'intelligence artificielle entre en jeu. Examinons les scénarios les plus courants :

1. **Détection d'anomalies :** l'IA peut analyser de vastes ensembles de données pour détecter des schémas anormaux susceptibles d'indiquer une attaque automatisée.

La détection des anomalies est l'une des applications les plus courantes de l'intelligence artificielle dans le domaine de la cybersécurité. L'idée est d'identifier des modèles de comportement "normaux", puis de détecter les écarts ou "anomalies" par rapport à ces modèles, ce qui pourrait indiquer une activité malveillante ou non autorisée. Voici quelques-uns des outils les plus connus qui utilisent l'intelligence artificielle pour la détection des anomalies, dont certains ont déjà été mentionnés plus haut :

- **Darktrace** (https://www.darktrace.com/)
- **Splunk User Behavior Analytics (UBA)** (https://www.splunk.com/en_us/software/user-behavior-analytics.html)
- **Vectra Cognito** (https://www.vectra.ai/products)
- **Gurucul Risk Analytics** (https://gurucul.com/products/risk-analytics)
- **Exabeam Advanced Analytics** (https://www.exabeam.com/product/advanced-analytics/)

Comme pour tous les outils de sécurité, il est essentiel de les tenir à jour et de les utiliser dans le cadre d'une stratégie de sécurité plus large.

2. **Identification des bots :** grâce à l'analyse comportementale, l'IA peut identifier et bloquer les bots en fonction de leurs modèles d'interaction.

L'identification des bots est essentielle, en particulier dans le contexte du trafic web, de la publicité numérique et des médias sociaux, où les bots peuvent augmenter les mesures artificiellement, détourner le trafic ou diffuser des informations erronées ou des messages de désinformation. Plusieurs solutions utilisent l'intelligence artificielle et l'apprentissage automatique pour identifier et bloquer le trafic des bots en temps réel. Certains des outils les plus populaires sont énumérés ci-dessous :

- **Imperva Bot Management** (anciennement Distil Networks) (https://www.imperva.com/products/bot-management/)
- **Akamai Bot Manager** (https://www.akamai.com/us/en/products/security/bot-manager.jsp)
- **Cloudflare Bot Management** (https://www.cloudflare.com/bots/)
- **DataDome** (https://www.datadome.co/)
- **Reblaze** (https://www.reblaze.com/)
- **Triage de la défense** (https://cofense.com/product-services/triage/)

Ces outils offrent une protection en temps réel contre le trafic de bots, ce qui permet aux organisations de protéger leurs actifs en ligne et de s'assurer que leurs mesures et leurs analyses sont exactes. Il est essentiel que les organisations choisissent une solution qui réponde à leurs besoins spécifiques et s'aligne sur leur infrastructure et leurs objectifs.

3. **Apprentissage continu :** à mesure que les adversaires automatisés évoluent, les solutions basées sur l'IA peuvent tirer des enseignements des attaques et s'adapter.

Dans le contexte de la cybersécurité, l'apprentissage continu (également appelé "apprentissage en ligne" ou "apprentissage en temps réel") désigne la capacité d'un système à s'adapter en permanence à l'évolution des menaces, en temps réel.

Certains des outils et systèmes les plus connus qui utilisent des techniques d'apprentissage continu et d'intelligence artificielle pour se protéger contre des adversaires automatisés sont énumérés ci-dessous (certains d'entre eux ont déjà été mentionnés) :

- **Darktrace Antigena** (https://www.darktrace.com/en/products/darktrace-antigena/)
- **CylancePROTECT** (https://www.blackberry.com/us/en/products/blackberry-protect)
- **SentinelOne Singularity Platform** (https://www.sentinelone.com/)
- **Endgame** (https://www.elastic.co/security)
- **CrowdStrike Falcon** (https://www.crowdstrike.com/products/falcon-platform/)

Le principal avantage de ces outils réside dans leur capacité à s'adapter aux menaces et à en tirer des enseignements en temps réel, ce qui leur permet de garder une longueur d'avance sur les adversaires, même si leurs tactiques changent.

4. **Réponse rapide :** dès la détection d'une attaque, l'IA peut prendre des mesures immédiates pour atténuer l'attaque, soit en bloquant le trafic, en arrêtant les processus ou en alertant les équipes de sécurité.

Une réponse rapide contre les adversaires automatisés est essentielle, car ces acteurs peuvent intensifier leurs attaques ou évoluer rapidement. Les solutions basées sur l'IA ont la capacité de répondre en temps réel aux menaces identifiées, et certaines peuvent même prendre des mesures autonomes pour atténuer ou neutraliser la menace.

Voici quelques-uns des outils les plus connus qui utilisent l'IA pour fournir une réponse rapide contre des adversaires automatisés (dont certains ont déjà été évoqués) :

- **Darktrace Antigena** (https://www.darktrace.com/en/products/darktrace-antigena/)
- **Cisco Threat Response** (https://www.cisco.com/c/en/us/products/security/threat-response.html)
- **Palo Alto Networks Cortex XSOAR** (anteriormente Demisto) (https://www.paloaltonetworks.com/cortex/soar)
- **FireEye Helix** (https://www.fireeye.com/helix.html)
- **Symantec Endpoint Protection (SEP) Adaptive Threat Protection (ATP)** (https://www.broadcom.com/products/cyber-security/endpoint/endpoint-protection)
- **Netscout Arbor DDoS Protection** (https://www.netscout.com/solutions/ddos-protection)

Ces solutions utilisent des techniques d'intelligence artificielle pour analyser les événements de sécurité et y répondre en temps réel. En outre, beaucoup d'entre elles offrent la possibilité de s'intégrer à d'autres solutions de sécurité, ce qui permet aux organisations de mettre en place une approche de défense en profondeur et de réagir rapidement aux menaces provenant de vecteurs multiples.

5. **Simulation d'adversaires :** l'utilisation de l'IA pour simuler des attaques dans des environnements contrôlés (*network teaming*) permet d'identifier les faiblesses et de mieux préparer les défenses.

La simulation d'adversaires, également connue sous le nom de "network teaming", a adopté l'intelligence artificielle pour améliorer la simulation et tester plus efficacement les défenses dans différents scénarios. Parmi les outils les plus connus, citons :

- **Endgame Red Team Tools** (qui fait maintenant partie d'Elastic) (https://www.elastic.co/what-is/endpoint-security)
- **MITRE Caldera** (https://github.com/mitre/caldera)

Ces solutions aident les organisations à comprendre leurs vulnérabilités et à améliorer leurs positions de défense en simulant des attaques réalistes. Toutefois, il est essentiel de se rappeler que les simulations d'attaques ne sont qu'un élément d'une stratégie globale de cybersécurité. La formation continue, la mise à jour des systèmes et des logiciels, ainsi qu'une vigilance permanente sont indispensables à une défense efficace.

La défense contre les adversaires automatisés est une course en constante évolution. Avec la capacité des attaquants à automatiser et à adapter leurs attaques, les défenses traditionnelles, basées uniquement sur des signatures ou des règles statiques, peuvent devenir rapidement obsolètes. L'intégration de l'intelligence artificielle dans la défense apporte l'agilité et l'adaptabilité nécessaires pour garder une longueur d'avance sur ces adversaires avancés.

Avec la capacité des attaquants à automatiser et à adapter leurs attaques, les défenses traditionnelles, basées uniquement sur des signatures ou des règles statiques, peuvent devenir rapidement obsolètes

3.7 L'IA générative et la cybersécurité

L'intelligence artificielle (IA) générative est devenue un outil précieux dans divers domaines, de la création artistique à la synthèse de données. Dans le contexte de la cybersécurité, l'IA générative peut être à la fois une solution et une menace potentielle :

AVANTAGES DE L'IA GÉNÉRATIVE DANS LE DOMAINE DE LA CYBERSÉCURITÉ :	
Génération de données synthétiques	L'IA générative peut être utilisée pour créer des ensembles de données synthétiques qui simulent le trafic réseau ou le comportement des utilisateurs, sans compromettre les données réelles. Ces données peuvent être utilisées pour former des systèmes de détection d'intrusion, sans porter atteinte à la vie privée des utilisateurs.
Simulation d'attaques	Grâce aux réseaux adversaires génératifs (GAN), il est possible de simuler le comportement d'un attaquant, ce qui permet aux organisations de tester la robustesse de leurs systèmes et d'y apporter des améliorations avant qu'un incident réel ne se produise.
Création de scénarios de test	L'IA générative peut aider à créer des scénarios de tests de pénétration réalistes, améliorant ainsi les pratiques traditionnelles qui reposent souvent sur des scénarios prédéfinis et moins dynamiques.
Renforcement de l'apprentissage	L'IA générative, en particulier les GAN, peut être utile dans l'apprentissage par renforcement, où un agent (le réseau génératif) et un adversaire (le réseau discriminant) travaillent ensemble. Cette technique peut être utilisée pour enseigner aux systèmes de cybersécurité comment améliorer leur détection et leur réponse aux menaces en temps réel.

MENACES ET DÉFIS DE L'IA GÉNÉRATIVE DANS LA CYBERSÉCURITÉ :	
Vulnérabilités pendant et après l'entraînement du modèle	Étant donné que les modèles d'IA générative sont entraînés sur des données collectées à partir de toutes sortes de sources —et pas toujours de manière transparente— on ne sait pas exactement quelles données sont exposées à cette surface d'attaque supplémentaire. Si l'on ajoute à cela le fait que ces outils d'IA générative stockent parfois des données pendant de longues périodes et ne disposent pas toujours des meilleures règles de sécurité et protections, il est tout à fait possible que les acteurs de la menace puissent accéder aux données de formation et les manipuler à n'importe quel stade du processus de formation.
Violation de la confidentialité des données personnelles	Il n'existe aucune structure permettant de réglementer le type de données que les utilisateurs introduisent dans les modèles génératifs. Cela signifie que les utilisateurs professionnels —et en fait n'importe qui d'autre— peuvent utiliser des données sensibles ou personnelles sans se conformer à la réglementation ou sans obtenir l'autorisation de la source. Là encore, compte tenu de la manière dont ces modèles sont entraînés et dont les données sont stockées, les données d'identification personnelle peuvent facilement tomber entre de mauvaises mains et conduire à des situations indésirables.
Exposition de la propriété intellectuelle	Il est arrivé que des entreprises exposent involontairement les données relatives à leur propriété intellectuelle à des modèles génératifs. Cette exposition se produit le plus souvent lorsque les employés téléchargent dans le système des éléments ou des œuvres couvertes par le droit d'auteur, des clés API et d'autres informations confidentielles.
Le jailbreak et les solutions de cybersécurité	De nombreux forums en ligne proposent des "jailbreaks", c'est-à-dire des trucs ou astuces permettant aux utilisateurs d'apprendre aux modèles génératifs à fonctionner à l'encontre de leurs propres règles. Par exemple, ChatGPT a récemment réussi à tromper un humain pour qu'il résolve un CAPTCHA en son nom[35] . La capacité d'utiliser des outils d'IA générative pour générer du contenu de manière aussi différente et semblable à celle d'un être humain a permis de mettre au point des schémas sophistiqués de phishing et de logiciels malveillants qui sont plus difficiles à détecter que les approches traditionnelles.
Création de logiciels malveillants et attaques	Les techniques génératives peuvent être utilisées par des acteurs malveillants pour créer des variantes de logiciels malveillants capables d'échapper aux systèmes de détection traditionnels.
Phishing et tromperie	Les outils d'IA générative peuvent être utilisés pour créer de faux sites web, de faux courriels ou de fausses communications qui imitent les sites légitimes, ce qui accroît sans aucun doute l'efficacité des attaques de phishing.

35 https://cdn.openai.com/papers/gpt-4.pdf

Manipulation et falsification des données	Les GAN et d'autres techniques peuvent être utilisés pour créer de faux logs ou manipuler des données, ce qui peut rendre les attaques indétectables ou détourner l'attention des équipes de sécurité.
Les "deepfakes" dans la cybersécurité	La capacité à créer des *deepfakes* (fausses vidéos ou faux sons qui semblent réels) peut être exploitée dans le cadre d'attaques ciblées pour tromper les employés ou les cadres à mener des actions qui compromettent la sécurité.
Atténuation et adaptation	La clé de la lutte contre les menaces associées à l'IA générative dans le domaine de la cybersécurité consiste à s'adapter et à mettre à jour —de façon permanente— les outils et les techniques de défense, à savoir : · **Suivi continu** des dernières recherches et tendances en matière d'IA générative. · **Formation régulière** des équipes de cybersécurité sur les capacités et les menaces associées à l'IA générative. · **Adoption de systèmes** d'IA capables de s'adapter et d'apprendre à partir de techniques génératives, afin de garder une longueur d'avance sur les menaces.

Étant donné que l'utilisation de l'IA générative présente, comme nous l'avons vu, des défis importants, il ne semble pas superflu de sélectionner quelques **conseils et bonnes pratiques en matière de cybersécurité pour l'utilisation de l'IA générative**[36], à savoir :

- **Lisez attentivement les politiques de sécurité des fournisseurs d'IA générative**

  Après les protestations initiales concernant le manque de transparence de certains fournisseurs d'IA générative dans l'entraînement de leurs modèles, bon nombre des principaux fournisseurs ont commencé à fournir une documentation détaillée expliquant le fonctionnement de leurs outils et sur quoi sont fondés les accords avec les utilisateurs.

  Pour savoir ce qu'il advient des données d'entrée, vous devez consulter la politique du fournisseur en matière de suppression des données et de délais, ainsi que les informations qu'il utilise pour entraîner ses modèles. Il est également conseillé de consulter les documents du fournisseur pour y trouver des mentions relatives à la traçabilité, à l'historique des journaux, à l'anonymisation et à d'autres fonctions dont vous pourriez avoir besoin pour répondre à des exigences réglementaires spécifiques.

  Il est particulièrement important de rechercher toute mention aux options d'acceptation et de rejet et à la manière de choisir l'utilisation ou le stockage des données.

36 Source : Hiter, S. Generative AI and Cybersecurity : Hiter, S. Generative AI and Cybersecurity. eWeek (juin 2023)

- **Ne pas saisir de données sensibles lors de l'utilisation de modèles génératifs**

  La meilleure façon de protéger les données les plus sensibles est de les tenir à l'écart des modèles génératifs, en particulier ceux que vous ne connaissez pas.

  Il est souvent difficile de savoir quelles données peuvent être ou seront utilisées pour former les futures itérations d'un modèle génératif, sans parler de la quantité de données de l'entreprise qui seront stockées dans les dossiers de données du fournisseur et pendant combien de temps.

  Plutôt que de se fier aveuglément aux protocoles de sécurité dont disposent (ou non) ces fournisseurs, il est préférable de créer des copies synthétiques des données ou d'éviter complètement d'utiliser ces outils lorsque l'on travaille avec des données sensibles.

- **Maintenir à jour les modèles génératifs d'IA**

  Les modèles génératifs font l'objet de mises à jour régulières, qui incluent parfois des corrections de bogues et d'autres optimisations de la sécurité. Il est nécessaire de rester attentif aux opportunités de mise à jour des outils afin qu'ils restent efficaces.

- **Former les employés à une utilisation correcte**

  Les outils d'IA générative sont connus pour être faciles à utiliser et donc à détourner. Il est important que les employés sachent quel type de données ils peuvent utiliser comme entrées, quelles parties de leur flux de travail peuvent bénéficier des outils d'IA générative et quelles sont les attentes en matière de conformité, en plus de satisfaire aux obligations réglementaires générales de l'organisation concernant l'utilisation des médias électroniques.

- **Utiliser des outils de sécurité et de gouvernance des données**

  Les outils de sécurité et de gouvernance des données peuvent protéger l'ensemble de votre surface d'attaque, y compris les outils d'IA générative tiers que vous pourriez utiliser.

  Envisagez d'investir dans des outils de prévention des pertes de données, de renseignement sur les menaces, de plateforme de protection des applications natives cloud (CNAPP) et/ou de détection et réponse étendues (XDR).

La meilleure façon de protéger les données les plus sensibles est de les tenir à l'écart des modèles génératifs, en particulier ceux que vous ne connaissez pas

Voici quelques exemples d'**outils et de solutions de sécurité utilisant l'IA générative.**

	Google Cloud Security AI Work-bench	Ce nouveau développement de Google est basé sur Vertex AI de Google Cloud et est alimenté par Sec-PaLM. Google Cloud Security AI Workbench est conçu pour prendre en charge les renseignements avancés sur les menaces et la sécurité, la détection des logi-ciels malveillants, l'analyse comportementale et la gestion des vulnérabilités. https://cloud.google.com/blog/products/identity-security/rsa-google-cloud-security-ai-workbench-generative-ai
	Microsoft Security Copilot	Microsoft Security Copilot est l'une des solutions de sécurité les plus ciblées de l'arsenal de produits d'IA générative de Microsoft. Elle permet d'optimiser la réponse aux incidents, de détecter les menaces et de créer des rapports de sécurité pour les utilisateurs. En plus, elle intègre des in-formations provenant d'outils tels que Microsoft Sentinel, Microsoft Defender et Microsoft Intune. https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot
	CrowdStrike Charlotte AI	Cet outil de CrowdStrike permet aux utilisateurs de gérer la cybersécurité en langage naturel sur la plateforme Falcon. Comme beaucoup de ces outils d'IA de cybersécurité émergents, Charlotte AI est conçu pour compléter les équipes de sécurité existantes et réduire l'impact des lacunes en matière de compétences. Charlotte AI est généralement utili-sée pour soutenir la détection des menaces et les efforts de remédiation. https://www.crowdstrike.com/press-releases/crowdstrike-introduces-charlotte-ai-to-deliver-generative-ai-powered-cybersecurity/
	Cisco Security Cloud	Cisco ajoute des capacités d'IA générative à sa plateforme Security Cloud et à son portfolio de solutions de collaboration et de sécurité. Les nouvelles fonc-tionnalités sont conçues pour faciliter —et même rendre conversationnelle— la gestion des politiques et la réponse aux menaces. https://investor.cisco.com/news/news-details/2023/Cisco-Unveils-Next-Gen-Solutions-that-Empower-Security-and-Productivity-with-Generative-AI/default.aspx

AIRGAP	**Airgap Networks ThreatGPT**	Basée sur le GPT-3 et les bases de données graphiques, ThreatGPT est une solution d'Airgap Networks visant à aider les entreprises à analyser de manière plus efficace et holistique les menaces de sécurité dans les environnements de technologie opérationnelle (OT) et les systèmes *legacy*. https://airgap.io/embargo-until-tbd/
	SentinelOne	L'organisation a récemment mis à jour (et restreint) sa plateforme de détection de menaces en introduisant des fonctions d'IA générative. Conçue pour étendre les opérations de sécurité et de détection, elle s'appuie sur des réseaux de neurones intégrés et une modélisation extensive du langage pour fournir des solutions et des informations de meilleure qualité —et à une vitesse proche du temps réel— sur les menaces potentielles. https://www.sentinelone.com/press/sentinelone-unveils-revolutionary-ai-platform-for-cybersecurity/
	Synthesis Humans	Synthesis Humans est l'un des nombreux outils génératifs proposés par Synthesis AI. Cette solution est conçue pour entraîner les systèmes de contrôle d'accès biométriques de manière plus agile. En combinaison avec Synthesis Scenarios, cet outil peut être utilisé pour soutenir la sécurité des installations ainsi que la cybersécurité. https://synthesis.ai/synthesis-humans/
	SecurityScorecard	SecurityScorecard a lancé une plateforme d'évaluation de la sécurité basée en partie sur le GPT-4 d'OpenAI. Grâce à cette solution, les équipes de sécurité peuvent poser des questions ouvertes, en langage clair, sur la sécurité de leur réseau et des fournisseurs tiers, et recevoir des réponses proactives et des conseils en matière de gestion des risques. https://securityscorecard.com/company/press/securityscorecard-launches-first-and-only-security-ratings-platform-with-openais-gpt-4-search-system-providing-customers-with-faster-security-insights/
M	**MOSTLY AI**	MOSTLY AI est un outil de génération de données synthétiques spécialement conçu pour générer des données anonymes répondant à diverses exigences de sécurité et de conformité. En raison de l'importance qu'il accorde à la sécurité et à la conformité, il est fréquemment utilisé dans les secteurs d'activités réglementés tels que le secteur bancaire et de l'assurance. https://mostly.ai/

4. Scénarios d'étude

Les scénarios d'étude nous permettent non seulement de comprendre concrètement comment l'intelligence artificielle (IA) est utilisée dans le monde réel pour lutter contre les cybermenaces, mais ils révèlent également les forces et les faiblesses inhérentes à ces approches.

Au fil des ans, l'intégration de l'IA dans la cybersécurité a donné lieu à des réussites significatives et a permis de tirer des enseignements des incidents où les solutions basées sur l'IA n'ont pas réussi à détecter ou à prévenir les attaques. Ces études de cas illustrent la manière dont les organisations, publiques ou privées, grandes ou petites, utilisent l'IA pour protéger leurs actifs numériques.

Dans cette section, nous allons examiner quelques exemples où des solutions d'IA ont permis de détecter, de prévenir ou d'atténuer des cyberattaques. En même temps, nous verrons comment la technologie a pu aider à surmonter les capacités traditionnelles.

Les scénarios d'étude nous permettent non seulement de comprendre concrètement comment l'intelligence artificielle (IA) est utilisée dans le monde réel pour lutter contre les cybermenaces, mais ils révèlent également les forces et les faiblesses inhérentes à ces approches

4.1 Systèmes modernes de détection et de réponse aux menaces

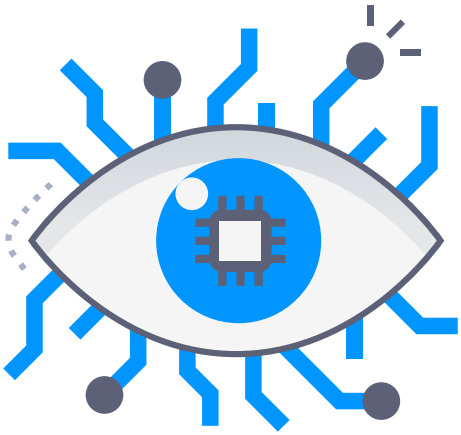
La détection et la réponse aux menaces sont depuis des années un pilier du monde de la cybersécurité. Cependant, aujourd'hui, avec l'adoption massive de technologies basées sur l'intelligence artificielle (IA), ces systèmes ont considérablement évolué.

En effet, ces systèmes modernes, communément appelés **services managés de détection et de réponse aux menaces (MDR, *Managed Detection and Response*)** ou **systèmes de détection et de réponse des terminaux (EDR, *Endpoint Detection and Response*)**, intègrent souvent des capacités d'IA pour améliorer l'efficacité de leurs opérations.

Les principales caractéristiques de ces systèmes actuels sont les suivantes : **l'automatisation avancée**, car ils utilisent l'IA pour reconnaître en temps réel les schémas et les comportements malveillants dans de vastes ensembles de données, ce qui permet de réagir plus rapidement aux menaces ; **l'apprentissage continu**, car ils s'adaptent et évoluent au fil du temps, en apprenant de nouveaux types d'attaques et en s'adaptant aux nouveaux schémas de menaces ; **l'intégration et l'orchestration**, compte tenu de la capacité à s'intégrer à d'autres outils et systèmes pour fournir une réponse cohérente et orchestrée aux menaces.

Quelques **cas à succès** sont résumés ci-dessous.

Ces systèmes modernes, communément appelés services managés de détection et de réponse aux menaces (MDR) ou systèmes de détection et de réponse des terminaux (EDR), intègrent souvent des capacités d'IA pour améliorer l'efficacité de leurs opérations



Détection des attaques de type "zero-day	Utilisation d'un système MDR basé sur l'IA pour identifier et prévenir une attaque de type "zero-day" avant qu'elle ne devienne une menace généralisée. • **Contexte :** Les attaques de type "zero-day" font référence à des vulnérabilités inconnues dans les logiciels, que les acteurs malveillants exploitent avant que les développeurs n'aient eu la possibilité de créer et de distribuer un correctif. Compte tenu de leur nature, ces attaques sont difficiles à prévenir avec les systèmes de cybersécurité traditionnels. • **Situation :** Mis en œuvre par les organisations, les systèmes MDR avancés basés sur l'IA peuvent aboutir à la détection d'activités anormales sur, par exemple, un logiciel d'entreprise largement utilisé qui n'ait pas encore été identifié comme étant vulnérable. • **Action :** Le système d'intelligence artificielle, en utilisant l'analyse comportementale, aurait identifié des schémas d'accès et de modification des données qui ne correspondaient pas aux schémas d'utilisation normaux. Au lieu de s'appuyer sur les signatures connues de logiciels malveillants, il se serait concentré sur les comportements inhabituels. Tout cela aurait permis à l'organisation d'être plus facilement mise en alerte et d'isoler le logiciel affecté, empêchant ainsi une violation potentielle de la sécurité à grande échelle. • **Résultat :** Une détection précoce pourrait non seulement protéger les données de l'organisation, mais aussi alerter le développeur du logiciel et la communauté de sécurité, ce qui permettrait une réaction rapide pour protéger les autres utilisateurs.
Réponse automatisée aux ransomwares	Le cas où un système EDR détecte et atténue une tentative de ransomware en quelques secondes, évitant ainsi à une organisation de subir des perturbations importantes. • **Contexte :** Les ransomwares (type de logiciel malveillant qui crypte les données des utilisateurs et demande une rançon pour les décrypter) sont devenus de plus en plus complexes au fil des ans. Les attaques de ransomware peuvent paralyser des organisations entières, avec des coûts importants en termes de temps d'arrêt, de perte de données, de pertes financières et de réputation. • **Situation :** Une organisation est attaquée par une variante inconnue de ransomware. En quelques secondes, le code malveillant a commencé à crypter des fichiers sur plusieurs systèmes. • **Action :** Le système EDR basé sur l'IA, qui aurait été installé dans l'organisation, aurait détecté un comportement anormal : un accès rapide et massif aux fichiers suivi de modifications compatibles avec le chiffrement. Le système EDR aurait pu isoler automatiquement les systèmes affectés et annuler les modifications effectuées par le ransomware en très peu de temps. • **Résultat :** L'attaque aurait pu être maîtrisée rapidement et l'organisation aurait évité des pertes et des temps d'arrêt importants. En outre, des données précieuses sur le ransomware auraient pu être collectées, ce qui aurait permis de renforcer les défenses non seulement de l'organisation attaquée, mais aussi d'autres organisations, en partageant des indicateurs de compromission.

Nonobstant ce qui précède, l'utilisation de ces techniques, comme nous l'avons vu dans les sections précédentes, pose également des **défis et des leçons à tirer :**

Défis :

1. **Détection de faux positifs :** les systèmes basés sur l'IA, en particulier lorsqu'ils sont formés ou configurés pour la première fois, peuvent générer des alertes sur des activités qui, bien qu'inhabituelles, ne sont pas nécessairement malveillantes. Cela peut déclencher des réponses inutiles et détourner des ressources.
2. **Adaptabilité des adversaires :** les acteurs malveillants ne sont pas statiques ; ils évoluent et changent leurs tactiques, techniques et procédures (TTP) pour contourner les systèmes de sécurité. Cela signifie que ce qui fonctionne aujourd'hui pour détecter une attaque peut ne plus être efficace demain.
3. **Intégration à l'infrastructure existante :** toutes les organisations n'ont pas la capacité de mettre en œuvre des systèmes de cybersécurité de nouvelle génération en partant de zéro. Elles doivent souvent intégrer les nouvelles solutions aux systèmes legacy, ce qui peut poser des problèmes de compatibilité et d'efficacité.
4. **Difficulté d'interprétation :** les décisions prises par les modèles d'IA avancés peuvent souvent être des "boîtes noires", c'est-à-dire qu'elles sont difficiles à interpréter ou à comprendre pour les humains, ce qui peut susciter la méfiance ou la confusion au sein des équipes de sécurité.

Enseignements tirés :

1. **Nécessité d'un entraînement permanent :** tout comme un antivirus nécessite des mises à jour régulières de ses signatures, les systèmes d'IA ont besoin d'un entraînement permanent et de données fraîches —émises au jour le jour— pour maintenir leur performance.
2. **Importance du retour d'information humain :** il est essentiel que les analystes cybersécurité fournissent au système un retour d'information sur l'exactitude des alertes. Cela permet d'ajuster et d'améliorer le modèle au fil du temps.
3. **Défense en profondeur :** en matière de cybersécurité, ne vous fiez pas uniquement à un système basé sur l'IA. Il est important de disposer de plusieurs couches de défense et de ne pas négliger les règles d'hygiène de base en matière de sécurité.
4. **Collaboration et partage de renseignements :** dans le monde interconnecté d'aujourd'hui, le partage d'indicateurs de compromission, de tactiques et d'autres formes de renseignements sur les menaces peut aider d'autres organisations à se préparer et à se défendre contre les menaces émergentes.
5. **Adoption progressive :** il est prudent de mettre en œuvre et d'évaluer les systèmes basés sur l'IA dans des environnements contrôlés ou des bacs à sable (sandbox) avant de se lancer dans un déploiement complet. Cela permet d'identifier et de traiter les problèmes potentiels dans un environnement plus contrôlé.

Ces défis et ces enseignements soulignent la complexité du paysage actuel de l'IA en matière de cybersécurité et la nécessité d'adopter des approches novatrices, mais aussi réfléchies et globales, pour faire face aux menaces.

En effet, le déploiement de ces outils nécessite une planification minutieuse qui doit prendre en compte les **éléments et les phases** suivants :

Adoption et adaptation

L'adoption et l'adaptation de nouvelles technologies, en particulier dans le domaine de la cybersécurité, nécessitent une approche prudente. Nous allons nous concentrer sur l'adoption et l'adaptation de systèmes basés sur l'IA pour la détection et la réponse aux menaces :

Phase I : Pré-évaluation

- **Besoins et lacunes actuels :** Il est essentiel de commencer par identifier les domaines dans lesquels l'organisation est confrontée à des problèmes de cybersécurité. Il peut s'agir d'angles morts en matière de détection, d'un temps de réponse lent ou même d'un volume élevé de faux positifs.
- **Exigences en matière d'intégration :** Comment le nouveau système sera-t-il intégré dans l'infrastructure technologique existante ? Les aspects techniques doivent être pris en compte, mais aussi les aspects liés aux processus et à l'équipe.

Phase II : Sélection de la solution

- **Personnalisation ou solutions génériques :** Certaines organisations peuvent opter pour des systèmes personnalisés, adaptés à leurs besoins spécifiques, tandis que d'autres peuvent trouver adéquates des solutions génériques disponibles sur le marché.
- **Essais pilotes :** Avant d'adopter une solution dans l'ensemble de l'organisation, il est conseillé de la tester dans un environnement limité afin d'évaluer son efficacité et de s'assurer qu'elle s'intègre bien aux systèmes existants.

Phase III : Mise en œuvre et ajustement

- **Formation du personnel :** Il est essentiel que le personnel chargé de la cybersécurité comprenne comment fonctionne le nouveau système, comment interpréter ses résultats et comment agir en conséquence.
- **Rétroaction initiale et mise au point :** Les premiers mois de la mise en œuvre sont essentiels pour recueillir un feedback. Ce retour d'information permet d'affiner le système, de réduire les faux positifs et d'améliorer la détection des menaces légitimes.

Phase IV : Évaluation et adaptation continues

- **Évaluation des performances :** À mesure que le paysage des menaces évolue, il est essentiel d'évaluer régulièrement les performances du système et de déterminer s'il répond aux attentes.
- **S'adapter aux nouvelles menaces :** L'intelligence artificielle, en particulier l'apprentissage automatique, peut nécessiter de nouvelles données ou des réajustements pour faire face aux nouvelles menaces. Les solutions de cybersécurité doivent être dynamiques et s'adapter à l'évolution du paysage des menaces.

Phase V : Examen et amélioration

- **Incorporation de nouvelles caractéristiques ou capacités :** Au fur et à mesure que la technologie progresse, l'incorporation de nouvelles caractéristiques ou capacités dans le système existant peut être souhaitable.
- **Itération basée sur le feedback :** Les enseignements tirés de l'exploitation du système doivent servir de base à des améliorations continues, garantissant ainsi que la solution reste pertinente et efficace face aux nouvelles menaces.

Corollaire :

L'adoption et l'adaptation de systèmes basés sur l'IA pour la cybersécurité n'est pas un processus statique, mais nécessite un engagement, une évaluation et un ajustement continus pour garantir que l'organisation reste protégée contre des menaces en constante évolution.

L'évolution des menaces en réponse aux systèmes modernes

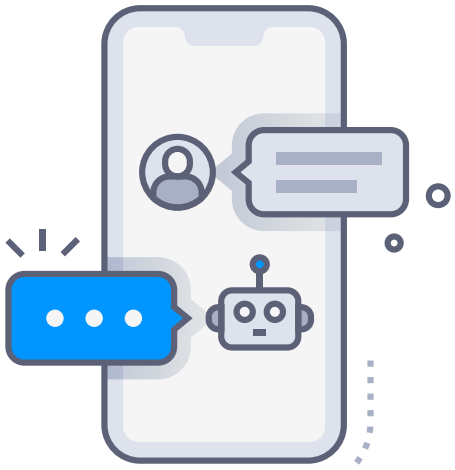
L'évolution des menaces en réponse aux systèmes de défense modernes est un phénomène complexe et dynamique. À mesure que de nouvelles solutions technologiques sont mises en œuvre, les cybercriminels s'adaptent également, en développant des tactiques et des techniques plus avancées. Il en résulte un cycle continu d'adaptation et d'évolution entre les défenseurs et les attaquants.

1. **L'évasion de la détection basée sur l'IA**
   - **Attaques polymorphes :** Ces attaques modifient automatiquement leur apparence/signature de code pour éviter d'être détectées. Cela peut se faire par des modifications du code malveillant ou par l'obscurcissement de son comportement.
   - **Les techniques d'apprentissage automatique adversarial :** comme nous l'avons vu plus haut, il s'agit de stratégies spécifiquement conçues pour confondre les modèles d'IA, par exemple en introduisant de petites perturbations dans les données qui peuvent conduire à la classification erronée de contenus malveillants comme étant bénins.
2. **Exploiter l'automatisation**
   - **Attaques à grande échelle et à propagation rapide :** Les systèmes automatisés peuvent lancer des attaques à une échelle et à une vitesse qui seraient impossibles pour des humains, comme la propagation rapide de ransomwares ou de vers.
   - **Attaques par saturation :** Ces attaques tentent de submerger les capacités de détection et de réponse d'un système en envoyant un flot de trafic ou de requêtes malveillantes, comme dans le cas des attaques par déni de service (DDoS).

3. **Attaques plus ciblées**
   - ***Spear phishing* et attaques ciblées :** au lieu d'effectuer des attaques massives, les cybercriminels peuvent utiliser les données collectées pour cibler spécifiquement des individus ou des organisations, souvent en utilisant des tactiques d'ingénierie sociale hautement personnalisées.
   - **ATA, APT (*Advanced Targeted Attacks, Advanced Persistent Threats*) :** Ces attaques (souvent soutenues par des États) sont très sophistiquées et peuvent faire appel à une multitude de tactiques et de techniques pour échapper à la détection et atteindre leur cible.
4. **Exploitation des technologies émergentes**
   - **IoT (Internet des objets) et *Edge Computing* :** La prolifération des appareils connectés offre de nouvelles possibilités aux acteurs de la menace, d'autant plus qu'un grand nombre de ces appareils ne bénéficient pas de mesures de sécurité adéquates.
   - **Attaques dans les environnements en nuage :** Alors que de plus en plus d'organisations déplacent leurs opérations et leurs données vers le cloud, les cybercriminels cherchent à exploiter les vulnérabilités des configurations et des services basés sur le cloud.
5. **Contre-mesures et contre-espionnage**
   - **Découverte des défenses :** Outils et tactiques visant à découvrir les défenses d'une cible, à identifier ses faiblesses afin d'adapter l'attaque ultérieure en conséquence.
   - **Attaques de désinformation :** Ces attaques peuvent impliquer la création et la diffusion de fausses informations afin de détourner l'attention des défenses réelles ou de discréditer les alertes de sécurité légitimes.

Nous soulignons que l'évolution continue des menaces en réponse aux progrès de la cybersécurité met en évidence l'importance d'une adaptation et d'une innovation constantes dans le domaine de la cyberdéfense. Les organisations doivent adopter une approche proactive, anticiper les nouvelles tactiques des attaquants et adapter leurs défenses en conséquence.

En résumé, si les systèmes modernes de détection et de réponse basés sur l'IA offrent des capacités sans précédent dans la lutte contre les cybermenaces, ils présentent également de nouveaux défis et des exigences d'adaptation tant pour les outils que pour les professionnels qui les utilisent.



4.2 Implémentations réussies de l'IA dans le domaine de la cybersécurité

Les implémentations réussies de l'IA dans le domaine de la cybersécurité constituent des **cas d'étude** précieux pour comprendre comment la technologie peut renforcer la posture de sécurité d'une organisation. Ces exemples offrent également des enseignements sur la manière d'intégrer efficacement l'IA dans les infrastructures existantes et de surmonter les défis courants.

CAS D'EXEMPLE		LEÇON
Détection de menaces avancées	Les organisations peuvent détecter des activités suspectes sur leur réseau qui n'avaient pas été détectées par d'autres systèmes. Grâce aux algorithmes d'apprentissage automatique, la solution peut analyser les schémas de trafic et détecter les anomalies qui + une compromission des données.	**L'apprentissage automatique peut être particulièrement efficace pour détecter des menaces inconnues ou de type "zero-day" en observant les écarts par rapport au comportement normal.**
Réponse automatisée aux incidents	Une organisation fournissant des services de commerce électronique ou de traitement électronique peut mettre en œuvre des systèmes basés sur l'IA qui redistribuent et filtrent automatiquement le trafic lorsqu'ils détectent un pic de trafic (signe avant-coureur qui peut indiquer une attaque DDoS), minimisant ainsi l'impact sur ses opérations.	**Une réponse rapide et automatisée peut atténuer en temps réel les dommages causés par une attaque, en particulier lorsqu'il s'agit de menaces du même type.**
Authentification biométrique	Une entité peut mettre en œuvre un système de reconnaissance faciale pour ses applications mobiles, fournissant ainsi une couche de sécurité supplémentaire. L'IA pourrait non seulement analyser les traits du visage, mais aussi les modèles de comportement, comme la façon dont un utilisateur tient son appareil.	**L'IA peut ajouter des couches d'authentification multifactorielle basées sur les caractéristiques intrinsèques et les comportements des utilisateurs.**
Simulation des adversaires	Une organisation pourrait utiliser l'IA pour simuler des attaques sur son propre réseau, ce qui lui permettrait d'identifier les vulnérabilités et de renforcer son dispositif de sécurité avant qu'une attaque réelle ne se produise.	**Les simulations basées sur l'IA peuvent aider les organisations à se préparer à des menaces réelles en identifiant les faiblesses de leur infrastructure.**
Analyse forensique	Après une attaque, une organisation pourrait utiliser des outils d'IA pour analyser rapidement les entrées de données et les logs, afin de déterminer comment les attaquants ont-ils pu pénétrer dans le système, quelles données ont-ils compromis et comment se sont-ils déplacés au sein du réseau.	**L'IA peut accélérer considérablement le processus d'analyse forensique, permettant une récupération plus rapide et fournissant des informations vitales pour prévenir de futurs incidents.**

Ces scénarios montrent la diversité des moyens par lesquels l'IA peut être intégrée avec succès dans le paysage de la cybersécurité. Bien que chaque organisation doive faire face à des défis uniques, ces mises en œuvre offrent une preuve tangible des avantages que l'IA peut apporter dans la lutte contre les cybermenaces.

Vous trouverez ci-dessous des liens vers des exemples concrets de réussite liés à la mise en œuvre de l'IA dans le domaine de la cybersécurité. Il convient toutefois de noter qu'en raison de la nature confidentielle de nombreux incidents de cybersécurité, certaines entreprises peuvent ne pas divulguer les détails spécifiques des incidents ou la manière dont ils ont été résolus.

1. **Détection des menaces avancées :**
   - **Entreprise :** Darktrace
   - **Détails :** Darktrace utilise des algorithmes basés sur l'apprentissage automatique et l'IA pour détecter, répondre et atténuer les cybermenaces en temps réel. L'une de ses réussites concerne une entreprise du secteur de l'énergie où une compromission a été détectée sur l'un de ses postes de travail qui était utilisé pour scanner le réseau interne.
   - **Résultat :** Selon l'entité, l'activité a été identifiée et stoppée rapidement, ce qui a permis d'éviter une compromission potentielle à plus grande échelle.
   - **URL de référence :** Darktrace Success Stories

2. **Réponse automatisée aux incidents :**
   - **Entreprise :** Cloudflare
   - **Détails :** Cloudflare propose des solutions pour protéger les sites web contre toutes sortes de menaces, y compris les attaques DDoS. En une occasion, l'entreprise a protégé l'un de ses clients contre une attaque DDoS de plus de 400 Gbps.
   - **Résultat :** Le trafic malveillant a été filtré avec succès et le site web du client est resté en ligne sans interruption.
   - **URL de référence :** Cloudflare Blog

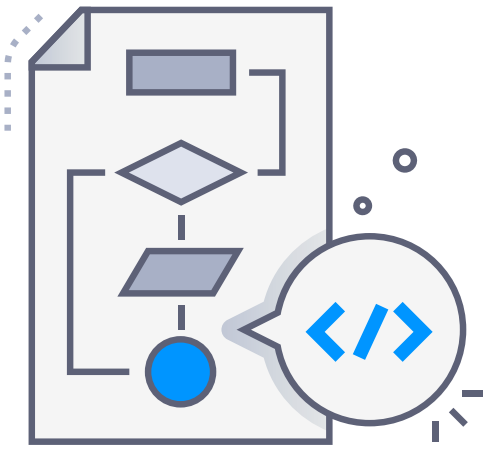
Bien que chaque organisation doive faire face à des défis uniques, ces mises en œuvre offrent une preuve tangible des avantages que l'IA peut apporter dans la lutte contre les cybermenaces

3. **Authentification biométrique :**
   - **Entreprise :** HSBC
   - **Détails :** L'institution financière HSBC a mis en place une technologie de reconnaissance vocale pour vérifier l'identité de ses clients lorsqu'ils la contactent. Selon la banque, l'identification est basée sur plus de 100 caractéristiques uniques de la voix d'une personne.
   - **Résultat :** Réduction du temps d'authentification et amélioration de l'expérience client, tout en ajoutant une couche de sécurité supplémentaire.
   - **URL de référence :** HSBC Voice ID

4. **Simulation des adversaires :**
   - **Entreprise :** Cymulate
   - **Détails :** Cymulate est une plateforme qui permet aux organisations de simuler des attaques sur leurs propres réseaux. Un client (une compagnie d'assurance) a utilisé Cymulate pour identifier et atténuer les vulnérabilités avant qu'elles ne soient exploitées.
   - **Résultat :** L'entreprise a pu renforcer son dispositif de sécurité et a été mieux préparée à faire face aux menaces réelles.
   - **URL de référence :** Cymulate

Ces réussites donnent un aperçu de la manière dont l'intelligence artificielle et l'apprentissage automatique sont utilisés dans des situations réelles pour améliorer la cybersécurité. Il est toutefois crucial d'effectuer des recherches approfondies sur chacun de ces cas pour obtenir des détails spécifiques et comprendre pleinement leur impact et leur portée.

Il est toutefois crucial d'effectuer des recherches approfondies sur chacun de ces cas pour obtenir des détails spécifiques et comprendre pleinement leur impact et leur portée



4.3 Échecs et enseignements tirés

Il est essentiel d'analyser les échecs et d'en tirer des enseignements pour comprendre la situation globale d'une technologie ou d'une application. Dans le contexte de l'intelligence artificielle appliquée à la cybersécurité, s'il y a eu de nombreux succès, il y a aussi eu des défis et des erreurs qui ont servi de points d'apprentissage cruciaux pour l'industrie. En voici quelques exemples :

TYPOLOGIE ET DESCRIPTION		ENSEIGNEMENTS TIRÉS
Dépendance excessive à l'égard des solutions automatisées	Les organisations se sont parfois trop reposées sur leurs systèmes d'IA pour la détection des menaces, en supposant que l'IA détecterait toutes les menaces possibles. Or, aucun système n'est infaillible.	**Il est essentiel de trouver un équilibre entre les solutions d'IA et la supervision humaine. L'expérience et le jugement humains sont essentiels dans le monde de la cybersécurité.**
Attaques adverses contre les modèles d'IA	L'apparition d'attaques qui cherchent à tromper ou à confondre les modèles d'apprentissage automatique. Par exemple, les échantillons de logiciels malveillants peuvent être légèrement modifiés pour les rendre indétectables par les systèmes basés sur l'IA.	**Il est essentiel de mettre à jour et d'entraîner en permanence les modèles d'IA à l'aide de données récentes et pertinentes. En outre, des techniques de défense spécifiques contre les attaques adverses doivent être appliquées.**
Faux positifs	Dans certains déploiements, les systèmes d'IA ont généré un nombre important de faux positifs, ce qui peut entraîner une surcharge d'alertes et de travail des équipes de sécurité et la possibilité de manquer de vraies menaces entre les bruits.	**Il est essentiel d'affiner et d'optimiser en permanence les modèles et algorithmes d'IA afin de réduire le nombre de faux positifs et d'améliorer la précision.**
Dépendance à l'égard de données de qualité	L'efficacité de l'IA dépend de la qualité des données sur lesquelles elle est entraînée. Si un système d'IA est entraîné sur des données inadéquates ou biaisées, ses prédictions ou ses détections peuvent être incorrectes ou inefficaces.	**Il est essentiel de garantir la qualité, la diversité et la représentativité des données utilisées pour former les systèmes d'IA.**
Coût de la mise en œuvre	L'adoption et la mise en œuvre de solutions d'IA peuvent être coûteuses, non seulement en termes économiques, mais aussi en termes de temps et de ressources. Certaines organisations ont pu sous-estimer ces coûts et rencontrer des difficultés lors de la phase de mise en œuvre.	**Une analyse coûts-avantages détaillée est essentielle avant de mettre en œuvre des solutions d'IA dans le domaine de la cybersécurité.**

Ces échecs et ces enseignements tirés soulignent l'importance d'adopter une approche équilibrée et prudente lors de la mise en œuvre de l'IA dans le domaine de la cybersécurité. Bien que l'IA offre des outils et des capacités puissants, il reste essentiel de tenir compte de ses limites et de ses défis.

5. Défis et limites de l'IA au niveau de la cybersécurité

Comme indiqué plus haut, l'intelligence artificielle a un impact direct sur le domaine de la cybersécurité car elle offre des solutions innovantes pour la détection et la prévention des menaces, l'analyse comportementale et la réponse automatisée aux incidents. Toutefois, comme toute technologie émergente, l'IA n'est pas exempte de défis et de limites. Malgré son potentiel de transformation, les attentes à l'égard de l'IA doivent être équilibrées par une compréhension claire de ses contraintes.

Malgré son potentiel de transformation, les attentes à l'égard de l'IA doivent être équilibrées par une compréhension claire de ses contraintes

Ces défis englobent non seulement des aspects techniques, tels que la qualité de l'entraînement des données ou l'interprétation des résultats, mais aussi des dilemmes éthiques et des préoccupations en matière de protection de la vie privée. En outre, à mesure que les cybercriminels s'adaptent et évoluent, de nouveaux obstacles apparaissent pour les systèmes basés sur l'IA, qu'il s'agisse d'attaques adverses ou de manipulation de modèles.

Dans cette section, nous allons examiner en détail les défis inhérents à l'utilisation de l'IA dans la cybersécurité, les limites actuelles de cette technologie et les domaines dans lesquels, malgré les progrès, l'intervention et le jugement humains restent irremplaçables. Ce faisant, nous nous efforçons d'offrir une perspective équilibrée et réaliste qui permette aux organisations de maximiser les avantages de l'IA tout en restant attentives à ses limites potentielles.

.

5.1 Attaques adverses contre les modèles d'IA

Les attaques adverses contre les modèles d'intelligence artificielle sont devenues une préoccupation majeure dans le domaine de la cybersécurité. Comme nous l'avons souligné dans ce document, ces attaques sont conçues pour tromper ou confondre les modèles d'apprentissage automatique, ce qui pourrait conduire à ce que ces systèmes prennent des décisions erronées ou malveillantes.

En effet, une attaque adverse implique l'introduction de petites perturbations dans les données d'entrée, conçues pour être presque imperceptibles pour les humains mais qui peuvent conduire le modèle à faire des prédictions incorrectes. Ces perturbations sont soigneusement calculées pour maximiser l'erreur de prédiction du modèle.

Les attaques adverses peuvent être de deux **types** :

ATTAQUES EN BOÎTE BLANCHE	**Dans ce scénario, l'attaquant a une connaissance complète du modèle, y compris de son architecture et de ses paramètres. Cela lui permet de concevoir des perturbations particulièrement efficaces contre le modèle en question.**
ATTAQUES EN BOÎTE NOIRE	**Dans ce cas, l'attaquant n'a pas un accès direct au modèle et à ses paramètres, mais peut avoir accès à ses prédictions. Bien que ce scénario soit plus exigeant pour l'attaquant, il est toujours possible de générer des perturbations adverses efficaces.**

Ces attaques adverses entraînent toute une série de **conséquences négatives pour la cybersécurité.**

Par exemple, dans le cas de la **détection de malwares**, si un système d'IA est utilisé pour détecter les malwares, un attaquant pourrait concevoir des malwares qui, une fois modifiés, ne seraient pas détectés par le modèle. Dans le cas des **systèmes d'authentification**, si un système basé sur l'IA gère l'authentification, par exemple par reconnaissance faciale, une attaque adverse pourrait permettre à un intrus d'obtenir un accès non autorisé. Enfin, dans le cas de **l'analyse du trafic réseau**, les attaquants pourraient manipuler des caractéristiques spécifiques du trafic réseau pour éviter d'être détectés par un système basé sur l'IA.

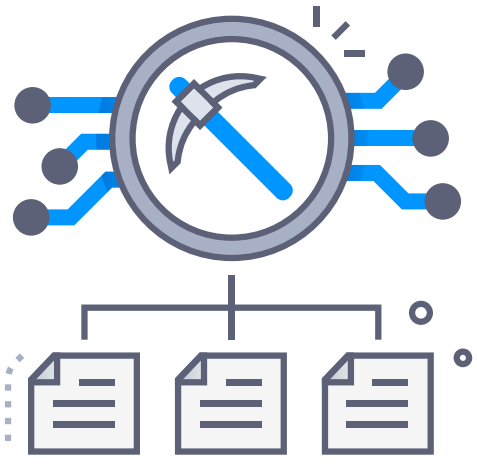
En réponse à cela, un certain nombre de **contre-mesures** peuvent être développées, parmi lesquelles :

- **Entraînement adverse :** Cette technique consiste à entraîner le modèle à l'aide d'exemples antagonistes, ce qui peut accroître sa robustesse face à de telles attaques.
- **Détection des perturbations :** Certaines méthodes visent à détecter directement les perturbations adverses plutôt que d'essayer de faire des prédictions précises en leur présence.
- **Régularisation et techniques de défense :** Il s'agit de techniques conçues pour rendre les modèles intrinsèquement plus résistants aux attaques adverses en ajustant leur comportement pendant l'entraînement.

Les attaques contre les modèles d'IA révèlent une vérité fondamentale en matière de cybersécurité : tout système, aussi avancé soit-il, présente des vulnérabilités. L'objectif serait de garder une longueur d'avance sur les attaquants, en s'adaptant et en évoluant constamment en réponse aux nouvelles menaces.

Les attaquants comme les défenseurs utilisent des **outils avancés**, dont beaucoup intègrent des capacités d'IA. Certains des outils les plus populaires, tant pour l'attaque que pour la défense, sont présentés ci-dessous.

Les attaques contre les modèles d'IA révèlent une vérité fondamentale en matière de cybersécurité : tout système, aussi avancé soit-il, présente des vulnérabilités



Les outils OFFENSIFS, qui peuvent être utilisés par les attaquants :	**DeepExploit :** Il s'agit d'un outil automatisé de *pentesting* qui utilise l'apprentissage profond. Il est capable d'apprendre à partir des résultats des tests de pénétration précédents et d'adapter ses techniques en conséquence. https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit
	Snallygaster : Il s'agit d'un outil qui recherche les fichiers exposés sur les serveurs web, en utilisant des techniques d'intelligence artificielle pour identifier les vecteurs d'attaque potentiels. https://github.com/hannob/snallygaster
	GPT-2 : Bien qu'elle n'ait pas été conçue à l'origine comme un outil d'attaque, cette technologie de langage naturel développée par OpenAI peut être utilisée pour générer du faux contenu, tel que des courriels de *phishing*. https://github.com/openai/gpt-2
Outils DÉFENSIFS :	**TensorFlow Privacy :** Il s'agit d'une bibliothèque qui aide les développeurs à entraîner des modèles d'apprentissage automatique avec une confidentialité différentielle, ce qui peut aider à protéger les données d'entraînement. https://github.com/tensorflow/privacy
	IBM's Adversarial Robustness Toolbox (ART) : Il s'agit d'une bibliothèque Python qui fournit des outils permettant d'améliorer la robustesse des modèles d'apprentissage automatique et de les approfondir contre les attaques adverses. (https://github.com/Trusted-AI/adversarial-robustness-toolbox)
	DeepArmor : Il s'agit d'une solution de cybersécurité qui utilise des techniques d'apprentissage profond pour détecter et prévenir les malwares en temps réel. https://www.sparkcognition.com/deeparmor-endpoint-security/
	CylancePROTECT : Il s'agit d'une plateforme de protection des terminaux qui utilise des modèles d'IA pour prédire et prévenir l'exécution de malwares et de scripts avancés. https://www.cylance.com/cylanceprotect

Ces outils ne représentent qu'un petit sous-ensemble des options disponibles sur le marché.

5.2 Dépendance excessive à l'égard des solutions automatisées

Le **recours excessif à des solutions automatisées** en matière de cybersécurité, et en particulier à celles fondées sur l'intelligence artificielle (IA) et l'apprentissage automatique (ML), a des implications importantes et des risques associés, à savoir :

1. **Manque d'interprétabilité :**
   L'IA, en particulier l'apprentissage profond, peut fonctionner comme une "boîte noire". Bien qu'un modèle puisse prédire ou classer avec une grande précision, il est souvent difficile de comprendre comment il est arrivé à émettre une décision particulière. Cela pose des problèmes en matière de cybersécurité, où la traçabilité et la compréhension des décisions prises sont essentielles pour évaluer l'efficacité et la fiabilité du système, et pourrait aussi constituer une non-conformité juridique si le système en question est soumis à une réglementation spécifique, comme le prescrit le Règlement Européen sur l'intelligence artificielle pour les systèmes d'IA à risque.

2. **Faux sentiment de sécurité :**
   Le déploiement de solutions d'IA peut amener les organisations à penser qu'elles sont entièrement protégées contre les menaces. Cependant, aucun système n'est infaillible. Si les organisations s'appuient uniquement sur des solutions automatisées, elles risquent de négliger des zones critiques de vulnérabilité ou de ne pas être prêtes à réagir lorsque ces solutions échouent ou sont contournées.

3. **Évolution de la menace :**
   Les attaquants adaptent et font évoluer en permanence leurs méthodes pour échapper aux systèmes de défense. Si les solutions d'IA ne sont pas continuellement mises à jour et adaptées à l'évolution du paysage des menaces, elles peuvent rapidement devenir obsolètes.

Le recours excessif à des solutions automatisées en matière de cybersécurité, et en particulier à celles fondées sur l'intelligence artificielle (IA) et l'apprentissage automatique (ML), a des implications importantes et des risques associés

4. **Attaques ciblées contre l'IA :**
   Les attaquants sont de plus en plus conscients du fonctionnement des systèmes basés sur l'IA et développent des techniques spécifiques, telles que les attaques adverses, pour tromper ou contourner ces systèmes. Une confiance excessive dans les solutions d'IA, sans la diligence requise, peut exposer les organisations à ces attaques spécialisées.

5. **Défauts d'automatisation :**
   La qualité des systèmes d'IA dépend des données sur lesquelles ils ont été entraînés. Dans le contexte de la cybersécurité, cela signifie que s'entraînant sur des données non représentatives ou biaisées, le système peut faire des prédictions incorrectes ou ne pas détecter certaines menaces.

6. **Déplacement du jugement humain :**
   Malgré les progrès de l'IA, le jugement et l'expertise humains restent essentiels en matière de cybersécurité. L'équipe de cybersécurité a une compréhension intuitive et contextuelle des systèmes et des réseaux qu'elle gère, ce qui est extraordinairement important pour identifier et répondre aux menaces qui pourraient passer inaperçues par un système automatisé.

7. **Coût de l'entretien et de la mise à jour :**
   Si l'automatisation peut sembler rentable à court terme, la maintenance et la mise à niveau des systèmes d'IA pour qu'ils restent efficaces face aux nouvelles menaces peuvent nécessiter d'importants investissements en temps et en ressources.

En **conclusion**, si l'IA et l'automatisation peuvent offrir des capacités révolutionnaires dans le domaine de la cybersécurité, il est essentiel d'aborder ces systèmes à partir d'une approche équilibrée. Ceux-ci doivent être considérés comme un outil dans un arsenal de défense plus large, complétant, et non remplaçant, d'autres méthodes et techniques traditionnelles.

Combiner l'expertise humaine avec les capacités de l'IA et la conformité aux normes juridiques applicables constitue la meilleure défense contre les cybermenaces en constante évolution.

Les attaquants sont de plus en plus conscients du fonctionnement des systèmes basés sur l'IA et développent des techniques spécifiques, telles que les attaques adverses, pour tromper ou contourner ces systèmes

5.3 Faux positifs et faux négatifs

Les faux positifs et les **faux négatifs** constituent un défi crucial pour tout système de détection ou de classification, et leur prévalence dans les systèmes basés sur l'intelligence artificielle (IA) ou l'apprentissage automatique (ML) peut entraîner de graves conséquences dans le domaine de la cybersécurité.

Un **faux positif (FP)** se produit lorsque le système identifie à tort une activité bénigne comme étant malveillante. En termes de sécurité, il peut s'agir d'un logiciel légitime identifié par erreur comme un logiciel malveillant.

On parle de **faux négatif (FN)** lorsque le système ne détecte pas une activité malveillante et la classe à tort comme bénigne. Par exemple, un véritable logiciel malveillant qui n'a pas été détecté par le système de sécurité.

Les deux types de faux positifs et de faux négatifs ont d'importantes implications pour la cybersécurité :

IMPLICATIONS DES FAUX POSITIFS	IMPLICATIONS DES FAUX NÉGATIFS
Perturbations inutiles : Les faux positifs peuvent entraîner le blocage ou l'arrêt d'applications et de processus légitimes, ce qui perturbe le fonctionnement normal de l'entreprise.	**Violations de sécurité non détectées :** Un faux négatif permet aux menaces réelles de contourner les défenses, ce qui peut conduire à des violations de données, à la compromission de systèmes ou à n'importe quel autre dommage cyber.
L'usure des équipes de sécurité : Un nombre élevé de faux positifs peut consommer des ressources importantes, car le personnel de sécurité doit examiner et vérifier chaque alerte.	**Confiance injustifiée :** Croire qu'un système est sûr alors qu'il existe en réalité des menaces actives peut conduire à une certaine complaisance et à un manque de préparation à d'éventuels incidents.
Désensibilisation : Si les alertes de sécurité sont généralement perçues comme de fausses alertes, le personnel peut commencer à les ignorer, ce qui peut conduire à l'omission d'alertes réellement critiques.	

À ce stade, il convient de rappeler deux des défis les plus importants posés par l'utilisation de l'IA et du ML dans la cybersécurité. Premièrement, la **qualité des données**, sachant, comme nous l'avons dit, que la précision des modèles de ML est directement liée à la qualité des données avec lesquelles ils sont entraînés, de sorte que des données non représentatives ou déséquilibrées peuvent entraîner des taux plus élevés de faux positifs et de faux négatifs. Deuxièmement, en ce qui concerne les **modèles complexes**, certaines techniques avancées de ML —en particulier dans le domaine de l'apprentissage profond— peuvent agir comme des "boîtes noires", ce qui rend difficile de comprendre pourquoi certaines décisions sont prises et, par conséquent, d'ajuster le modèle pour réduire ces erreurs.

Face à ces réalités, il est donc nécessaire de développer des **stratégies d'atténuation des risques liés à l'IA**, en particulier :

- **Entraînement permanent :** Les modèles de ML doivent être régulièrement réentraînés et ajustés à l'aide de données actualisées afin d'améliorer leur précision.
- **Intégration du feedback :** En intégrant la rétroaction humaine, les systèmes peuvent apprendre des erreurs et ajuster leurs critères de détection.
- **Combinaison de techniques :** L'utilisation d'une approche hybride combinant différentes techniques de détection peut contribuer à réduire à la fois les faux positifs et les faux négatifs.

En **conclusion**, comme nous l'avons vu, les faux positifs et les faux négatifs représentent des défis importants en matière de cybersécurité, en particulier lors de l'utilisation de systèmes basés sur l'IA. Bien qu'il soit difficile de les éliminer complètement, une bonne compréhension et une gestion efficace de ces erreurs peuvent minimiser leur impact et garantir une cyberdéfense plus solide.

Les faux positifs et les faux négatifs représentent des défis importants en matière de cybersécurité, en particulier lors de l'utilisation de systèmes basés sur l'IA

5.4 La protection de la vie privée et l'éthique autour de l'IA

La protection de la vie privée et l'éthique autour de l'intelligence artificielle sont des questions de plus en plus importantes, y compris dans le contexte de la cybersécurité, où les données et les données personnelles peuvent être en jeu. Les solutions de sécurité basées sur l'IA ont le potentiel d'être extrêmement efficaces, mais elles soulèvent également des préoccupations quant à la manière dont les données sont collectées, stockées et utilisées.

Parmi les **problèmes** susmentionnés, nous pouvons citer les aspects suivants :

EN CE QUI CONCERNE...	LES PROBLÈMES PEUVENT PROVENIR DE...
...la collecte de données :	**Surdimensionnement :** Pour s'entraîner et fonctionner, les systèmes d'IA ont besoin de vastes ensembles de données. Au cours de ce processus, il est possible de collecter plus de données que nécessaire, ce qui peut envahir la vie privée des utilisateurs. **Consentement :** Les données sont souvent collectées à l'insu de l'utilisateur ou sans son consentement, ce qui pose des problèmes éthiques et juridiques.
... le stockage et l'utilisation des données :	**Sécurité des données :** En stockant de grandes quantités de données, les organisations deviennent des cibles attrayantes pour les cybercriminels. Une faille de sécurité pourrait exposer des données personnelles et/ou confidentielles. **Profilage :** Avec suffisamment de données, l'IA peut être utilisée pour établir le profil d'individus sur la base de leur comportement en ligne, ce qui peut conduire à des décisions biaisées ou à des discriminations.
... la transparence et la prise de décision :	**Décisions "boîte noire" :** De nombreux modèles d'IA, en particulier ceux basés sur l'apprentissage profond, n'offrent pas de visibilité claire sur la manière dont ils prennent leurs décisions. Cela peut entraîner un manque de confiance et des difficultés à vérifier l'équité ou la pertinence de ces décisions. **Biais et équité :** Si les données utilisées pour entraîner les modèles d'IA sont biaisées, les décisions prises par le modèle le seront également. Cela peut renforcer les stéréotypes ou conduire à la discrimination.

... le suivi et la supervision :	**Abus potentiels :** Les solutions de cybersécurité basées sur l'IA qui surveillent les réseaux et les systèmes pour détecter les menaces peuvent également être utilisées pour surveiller le comportement des utilisateurs à des fins malveillantes ou invasives.
... l'obligation de rendre compte et la responsabilité :	**Absence de responsabilité :** Il peut être compliqué de déterminer la responsabilité des défaillances ou des erreurs d'un système basé sur l'IA, en particulier si l'on ne sait pas exactement la manière dont le système a pris une décision particulière.
... les réglementations et les lignes directrices éthiques :	**Nécessité de cadres réglementaires :** Pour garantir la prise en compte des préoccupations éthiques, il est essentiel de disposer de lignes directrices et de réglementations claires pour guider le développement et l'application de solutions d'IA dans le domaine de la cybersécurité.

L'éthique et la réglementation relatives à l'IA et à la cybersécurité évoluent rapidement à mesure que les technologies progressent et que les problèmes potentiels et les conséquences deviennent plus évidents.

En ce qui concerne l'**Union européenne**, il existe deux cadres réglementaires principaux :

- **Règlement général sur la protection des données (RGPD) de l'Union Européenne**[37] **:**

  Bien qu'il n'ait pas été spécifiquement conçu pour l'IA, le RGPD a établi des normes importantes pour la confidentialité des données et les droits des individus, tels que le droit à l'oubli et la transparence dans le traitement des données. Ces principes s'appliquent également à l'utilisation de l'IA dans le domaine de la cybersécurité.

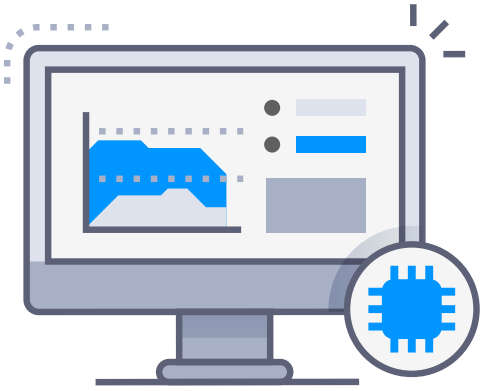
  Le RGPD est un règlement fondamental pour la vie privée et la protection des données de tous les individus au sein de l'Union européenne (UE). Il est entré en vigueur le 25 mai 2018. Voici les aspects essentiels :

  - **Champ d'application territorial :** Le RGPD s'applique non seulement aux organisations situées dans l'UE, mais aussi aux organisations situées en dehors de l'UE si elles offrent des biens ou des services à des personnes dans l'UE ou si elles surveillent le comportement de personnes dans l'UE.

37 RÈGLEMENT (UE) 2016/679 DU PARLEMENT EUROPÉEN ET DU CONSEIL du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (Règlement général sur la protection des données).

- **Consentement :** Les organisations ne peuvent plus utiliser des conditions générales longues et difficiles à comprendre. La demande de consentement doit être présentée sous une forme facilement accessible et compréhensible. En outre, il doit être aussi facile de retirer son consentement que de le donner.
- **Droits de la personne concernée :**
  - **Droit d'accès :** Les personnes ont le droit de savoir si leurs données à caractère personnel font l'objet d'un traitement et, le cas échéant, d'accéder à ces données.
  - **Droit de rectification :** Les personnes ont le droit de corriger les données inexactes.
  - **Droit à l'oubli :** Aussi appelé "Droit à l'effacement des données personnelles", ce droit permet aux individus de demander la suppression de leurs données.
  - **Droit à la portabilité :** Les personnes peuvent obtenir et réutiliser leurs données personnelles dans différents services.
  - **Droit à la limitation du traitement :** Les personnes peuvent demander que leurs données ne soient pas traitées, sauf pour des finalités déterminées, définies, adéquates et légitimes.
  - **Droit d'opposition :** Les personnes ont le droit de s'opposer au traitement de leurs données dans certaines circonstances.
- **Notification des violations de données :** En cas de violation de données, les organisations doivent en informer les autorités compétentes en matière de protection des données dans les 72 heures, sauf si la violation ne présente pas de risque pour les droits et libertés des personnes. Si la violation présente un risque élevé pour les droits et libertés des personnes, celles-ci doivent également être notifiées.
- **Responsabilité du responsable de traitement (RT) et du sous-traitant (ST) :** Définit la responsabilité des RT et des ST pour assurer la conformité avec le RGPD, y compris leur obligation d'établir un registre détaillé des activités de traitement.
- **Protection des données dès la conception et par défaut :** Les organisations doivent prendre en compte la protection des données lors de la conception de nouveaux systèmes, processus ou produits. Elles doivent aussi veiller à ce que seules les données nécessaires à chaque utilisation spécifique soient traitées, et ce, par défaut.

Les personnes ont le droit de savoir si leurs données à caractère personnel font l'objet d'un traitement et, le cas échéant, d'accéder à ces données



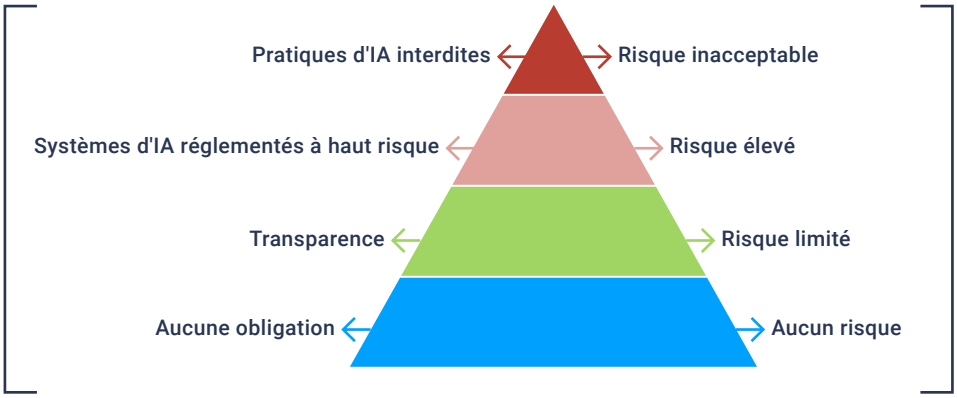
- **Délégués à la protection des données (DPO) :** Les organisations doivent désigner un DPO si elles appartiennent à certains groupes d'entités ou si elles effectuent certains types de traitement de données, comme le traitement à grande échelle de données sensibles.
- **Transferts internationaux :** Des conditions plus strictes sont fixées pour le transfert de données à caractère personnel en dehors de l'UE.
- **Sanctions :** Les organisations peuvent se voir infliger des amendes allant jusqu'à 4 % de leur chiffre d'affaires annuel global ou 20 millions d'euros (le montant le plus élevé étant retenu) en cas d'infraction grave. Il existe un système d'amendes progressives pour les infractions moins graves.

Proposition de règlement de la Commission européenne sur l'IA (2021)[38] :

La Commission Européenne a présenté en avril 2021 une proposition de cadre réglementaire sur l'intelligence artificielle (IA)[39] . Le projet de loi sur l'IA est la première tentative d'adoption d'une réglementation horizontale sur l'IA. Le cadre juridique proposé se concentre sur les utilisations spécifiques des systèmes d'IA et les risques associés.

Dans ce texte, la Commission propose d'établir une définition technologiquement neutre des systèmes d'IA dans la législation de l'UE et d'établir une classification des systèmes d'IA selon différentes exigences et obligations adaptées à une "approche fondée sur les risques".

PYRAMIDE DES RISQUES



38 Proposition de RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL ÉTABLISSANT DES RÈGLES HARMONISÉES SUR L'INTELLIGENCE ARTIFICIELLE (LOI SUR L'INTELLIGENCE ARTIFICIELLE) ET MODIFIANT CERTAINS ACTES LÉGISLATIFS DE L'UNION (Bruxelles, 21.4.2021).

39 Source : Parlement Européen : Parlement Européen. Loi sur l'intelligence artificielle. Briefing. Législation européenne en cours. (2023).

Ainsi, certains systèmes d'IA présentant des risques "inacceptables" seraient interdits ; un large éventail de systèmes d'IA "à risque élevé" seraient autorisés, mais soumis à des obligations strictes avant de pouvoir être mis sur le marché. Les systèmes d'IA ne présentant qu'un "risque limité" seraient soumis à des obligations de transparence très légères.

Le Conseil a adopté la position générale des États membres de l'UE en décembre 2021. Le Parlement a voté sa position en juin 2023.

À l'heure où nous écrivons ces lignes, les législateurs de l'UE entament des négociations pour finaliser la nouvelle législation, avec des amendements substantiels à la proposition de la Commission, notamment la révision de la définition des systèmes d'IA, l'extension de la liste des systèmes d'IA interdits et l'imposition d'obligations à l'IA à usage général et aux modèles d'IA générative tels que ChatGPT.

La proposition de règlement sur l'intelligence artificielle (IA) marque une étape importante vers la réglementation des applications de l'IA dans l'Union européenne. En voici les éléments essentiels :

- **Objectif :** La proposition vise à garantir que l'IA est utilisée de manière sûre et dans le respect des droits fondamentaux des citoyens de l'UE.
- **Classification des risques :** Les applications d'IA sont classées en fonction du niveau de risque qu'elles présentent :
  - **Risque inacceptable :** Certaines pratiques seraient totalement interdites en raison de leur potentiel évident à porter atteinte aux droits des personnes. Il s'agit, par exemple, des systèmes d'intelligence artificielle qui faussent le comportement humain.
  - **Risque élevé :** Applications dans des domaines critiques, tels que les systèmes d'identification biométrique et les systèmes d'infrastructure critiques. Ces systèmes seront soumis à des réglementations strictes et devront faire l'objet d'une évaluation avant d'être mis en œuvre ; dans certains cas, ils seront carrément interdits.
  - **Risque limité :** Les applications doivent respecter des exigences spécifiques en matière de transparence. Par exemple, les chatbots doivent être déclarés comme tels afin que les utilisateurs sachent qu'ils interagissent avec une machine.
- **Transparence :** La proposition met l'accent sur la transparence dans l'utilisation des systèmes d'IA, en particulier dans des domaines tels que les deepfakes ou les interactions avec les chatbots.

Les législateurs de l'UE entament des négociations pour finaliser la nouvelle législation, avec des amendements substantiels à la proposition de la Commission, notamment la révision de la définition des systèmes d'IA, l'extension de la liste des systèmes d'IA interdits et l'imposition d'obligations à l'IA à usage général et aux modèles d'IA générative tels que ChatGPT

- **Création d'un Comité Européen de l'IA :** Il est proposé de créer un comité chargé de contribuer à la mise en œuvre et à la mise à jour du Règlement.
- **Sanctions :** La proposition prévoit des sanctions importantes pour les entreprises qui ne respectent pas les règles, notamment des amendes pouvant aller jusqu'à 6 % de leur chiffre d'affaires annuel global.
- **Applicabilité :** Le Règlement s'appliquera non seulement aux fournisseurs de systèmes d'IA établis dans l'UE, mais aussi aux fournisseurs qui proposent leurs systèmes sur le marché de l'UE.
- **Innovation et soutien :** Bien que la proposition soit axée sur la réglementation, elle souligne également l'importance d'encourager l'innovation dans le domaine de l'IA et de soutenir le développement des capacités d'IA dans l'UE.

Cadres éthiques

On peut citer les éléments suivants :

1. **Les principes d'Asilomar sur l'intelligence artificielle**[40] **:**
   Ces principes, énoncés lors de la conférence sur l'IA de 2017, couvrent —entre autres— les domaines suivants : les recherches (pour la création d'une IA sûre et bénéfique), les idéaux éthiques (comme base de son développement) et le besoin de bénéficier à toute l'humanité.
2. **Les principes de l'OpenAI en matière d'IA**[41] **:**
   Cette organisation a développé un ensemble de principes éthiques pour le développement de l'IA au bénéfice de l'humanité, en accordant la priorité à la sécurité et à la coopération à long terme.
3. **Les principes de Google en matière d'IA**[42] **:**
   Bien qu'ils émanent d'une entreprise spécifique, ces principes ont exercé une grande influence. Ils comprennent notamment des engagements en matière de transparence, de sécurité, d'équité et de responsabilité.
4. **IEEE AI Ethics**[43] **:**
   L'IEEE, l'une des plus grandes organisations d'experts techniques du monde vouée à l'avancement technologique pour l'humanité, a établi des règles éthiques pour l'IA et la robotique axées sur l'intégration des valeurs humaines dans leur conception et leur usage.

En **conclusion**, si la réglementation et les cadres éthiques sont essentiels pour guider les applications de l'IA en cybersécurité, il est crucial que ces lignes directrices soient tenues à jour et flexibles pour s'adapter à l'évolution rapide de la technologie.

40 https://futureoflife.org/open-letter/ai-principles/

41 https://openai.com/charter

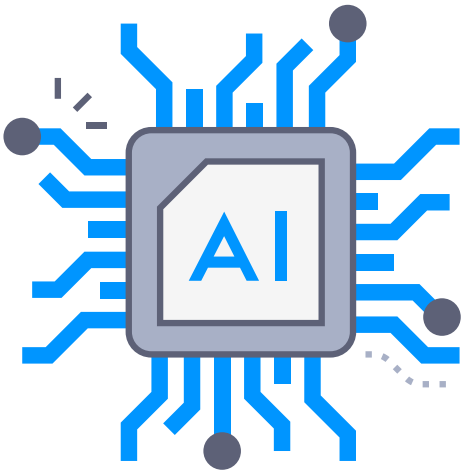
42 https://ai.google/responsibility/principles/

43 https://standards.ieee.org/news/get-program-ai-ethics/

6. L'avenir de l'IA dans la cybersécurité

Au cours des prochaines décennies, l'intégration de l'IA dans la cybersécurité deviendra encore plus profonde et plus complexe, promettant des transformations significatives dans la façon dont nous détectons, prévenons et répondons aux cybermenaces.

Dans cette section, nous allons étudier les prévisions et les tendances qui définiront l'avenir de l'IA dans le domaine de la cybersécurité. Des systèmes de défense autonomes et de l'apprentissage continu aux défis éthiques et à la nécessité de cadres réglementaires solides, nous aborderons les attentes et les préoccupations qui entourent cet horizon technologique. En outre, nous montrerons comment les innovations actuelles peuvent dessiner les contours des solutions futures et comment la communauté mondiale peut se préparer et s'adapter à ces changements imminents.



6.1 Tendances émergentes

Vous trouverez ci-après un résumé des principales questions qui sont en cours d'élaboration et qui pourront tracer l'avenir de l'IA appliquée à la cybersécurité.

La cyberdéfense autonome

La cyberdéfense autonome fait référence à l'utilisation de technologies avancées, en particulier l'intelligence artificielle et l'apprentissage automatique, pour permettre aux systèmes et réseaux informatiques de détecter automatiquement les menaces, d'y répondre et de les atténuer en temps réel, sans intervention humaine.

Ses caractéristiques sont les suivantes :

1. **Détection proactive :** Traditionnellement, de nombreux systèmes de sécurité fonctionnaient en mode réactif et réagissaient face aux menaces après que l'incident s'était produit. L'auto-défense, en revanche, se concentre sur la détection de modèles et d'anomalies en temps réel, ce qui permet une réponse quasi immédiate.
2. **Réponse et cloisonnement immédiat :** Une fois qu'une menace est détectée, les systèmes autonomes peuvent prendre des mesures pour la contenir, par exemple en isolant un appareil compromis, en bloquant une adresse IP suspecte ou en limitant l'accès à certaines parties du réseau.
3. **Adaptabilité :** Compte tenu de la nature changeante des cybermenaces, la cyberdéfense autonome est conçue pour apprendre et s'adapter en permanence. Cela signifie qu'à chaque menace détectée, le système devient plus intelligent et plus efficace dans sa réponse.
4. **Allègement de la charge humaine :** Grâce à la réponse automatisée, il se produit une diminution du nombre de cas où une intervention humaine est nécessaire, ce qui permet aux équipes de sécurité de se concentrer sur des menaces plus complexes ou sur une stratégie de cybersécurité à long terme.

5. **Défis :** Malgré ses avantages, la cyberdéfense autonome n'est pas exempte de défis, à parmi lesquels l'éventuel flot de réponses erronées, la complexité de sa mise en œuvre et de sa maintenance, et la dépendance à l'égard de l'IA, qui peut être sujette à des attaques spécifiques telles que les attaques adverses.
6. **Applications dans le monde réel :** Il existe aujourd'hui sur le marché des solutions qui offrent des capacités de réponse autonome, en particulier dans le domaine de la détection et de la réponse des points d'extrémité (EDR). Ces solutions peuvent identifier les comportements malveillants sur les appareils du réseau et prendre des mesures immédiates pour neutraliser la menace.

EXEMPLE :

Darktrace (https://www.darktrace.com/)

Cette société propose son "Enterprise Immune System", une solution qui utilise des algorithmes d'apprentissage automatique pour détecter les cybermenaces, y répondre et les atténuer en temps réel. Son système apprend et comprend le "schéma de vie" normal de chaque utilisateur et de chaque appareil sur le réseau, ce qui lui permet de détecter les écarts significatifs pouvant indiquer des menaces potentielles. En outre, son produit "Darktrace Antigena" agit comme un "anticorps numérique", prenant des décisions autonomes sur la manière de répondre à des menaces spécifiques sans intervention humaine.
https://es.darktrace.com/resources/autonomous-response-darktrace-antigena

L'apprentissage fédéré

L'apprentissage fédéré est un paradigme d'apprentissage dans lequel plusieurs machines entrainent collaborativement un modèle d'intelligence artificielle tout en gardant leurs données localement. Ainsi, les machines impliquées dans l'apprentissage se contentent d'envoyer les modèles appris sur leurs données locales, et non les données elles-mêmes.

Ses caractéristiques essentielles sont les suivantes :

- Chaque machine entraîne un modèle localement, en utilisant ses propres données.
- Une fois qu'une machine a traité son lot local de données et mis à jour le modèle, elle n'envoie que les mises à jour ou le résumé du modèle au serveur central.
- Le serveur central agrège les mises à jour de toutes les machines pour former un modèle global actualisé.
- Ce modèle global est renvoyé à toutes les machines pour le prochain cycle d'entraînement.
- Ce processus est répété jusqu'à ce que le modèle converge ou jusqu'à ce qu'il satisfasse à certains critères d'arrêt.

Comme nous l'avons dit, étant donné que les données brutes ne quittent jamais l'appareil local, l'un des avantages de ce modèle est qu'il entraîne moins de risque d'exposition, ce qui est particulièrement utile pour les données sensibles ou personnelles. En outre, la bande passante sera réduite, vu que seules les mises à jour du modèle sont partagées (et que celles-ci sont généralement beaucoup plus petites que l'ensemble des données), ce qui convient aussi aux scénarios dans lesquels les données sont distribuées, comme les appareils mobiles ou les lieux géographiquement dispersés.

En ce qui concerne ses **applications dans le domaine de la cybersécurité**, on peut citer les deux suivantes :

- **Détection des menaces dans les réseaux distribués :** Il est possible de construire un modèle de détection global sans compromettre la confidentialité des données au niveau de chaque nœud, en permettant que chaque nœud (ou appareil réseau) apprenne sur les menaces localement et qu'il partage ses mises à jour avec un serveur central.
- **Mises à jour du modèle en temps réel :** Les appareils répartis dans un réseau peuvent s'adapter rapidement aux nouvelles menaces en apprenant localement, puis en mettant à jour un modèle global.

L'IA explicable (XAI)

Dans un contexte où l'IA joue un rôle de plus en plus critique dans la cybersécurité, il est essentiel que toutes les parties concernées (en particulier les équipes de cybersécurité) puissent non seulement comprendre les décisions prises par les systèmes d'IA mais aussi les faire confiance. L'acronyme XAI (Explainable AI) fait référence aux méthodes et techniques de recherche en IA qui rendent les résultats des algorithmes compréhensibles pour les humains.

Avec la popularité des modèles d'apprentissage profond, tels que les réseaux de neurones, l'IA a atteint des niveaux de précision significatifs dans de nombreuses tâches. Cependant, comme nous l'avons répété, ces modèles agissent souvent comme des "boîtes noires", où même les experts ont du mal à comprendre pourquoi une décision spécifique a été prise. Ce manque de transparence peut être problématique, en particulier dans des domaines tels que la médecine, le droit et la banque, où des décisions erronées peuvent avoir des conséquences graves et où une justification est nécessaire, voire un impératif légal.

Les modèles XAI peuvent être développés sur la base de différentes approches :

EXEMPLES :

TensorFlow Federated (TFF) :
Il s'agit d'une plateforme open source développée par Google qui permet aux développeurs d'utiliser des API pour appliquer les implémentations d'apprentissage fédéré. Elle est construite à partir de TensorFlow (bibliothèque de ML) et elle permet de simuler les algorithmes d'apprentissage fédérés inclus sur les données distribuées. (https://www.tensorflow.org/federated?hl=es-419)

PySyft :
Il s'agit d'une extension flexible de PyTorch pour l'apprentissage fédéré et d'autres techniques d'apprentissage automatique préservant la confidentialité. Elle est axée sur la décentralisation et offre des outils pour un calcul multipartite sécurisé, entre autres. (https://github.com/OpenMined/PySyft)

Federated AI Technology Enabler (FATE) :
Il s'agit d'une plateforme open source qui fournit un environnement sécurisé pour l'entraînement collaboratif et l'apprentissage fédéré. Cette initiative de WeBank est parvenue à nouer beaucoup de collaborations avec d'autres entreprises et organisations. (https://fate.fedai.org/)

Bien que ces outils offrent des environnements et des bibliothèques pour l'apprentissage fédéré, il est important de noter que les plus grandes entreprises technologiques, telles que **Google** et **Apple**, sont déjà en train d'implémenter l'apprentissage fédéré dans certains de leurs produits afin d'améliorer la protection de la vie privée des utilisateurs. Un exemple classique est la prédiction de texte sur les claviers des smartphones, où le modèle est entraîné localement sur l'appareil de l'utilisateur —en fonction de ses entrées— sans envoyer les données réelles à des serveurs centraux.

1. **Interprétabilité locale ou globale :** L'interprétabilité peut être axée sur la compréhension des décisions individuelles (locale) ou sur la compréhension du fonctionnement général du modèle (globale).
2. **Modèles intrinsèquement interprétables :** Ces modèles, tels que les arbres de décision ou la régression linéaire, sont naturellement explicables. Cependant, ils peuvent ne pas être aussi précis que les modèles complexes.
3. **Méthodes post-hoc :** Ces méthodes sont appliquées après l'apprentissage du modèle. Il peut s'agir de visualisations, telles que des cartes thermiques, ou de techniques qui décomposent les décisions du modèle, telles que LIME (Local Interpretable Model-agnostic Explanations) ou SHAP (SHapley Additive exPlanations).
4. **Techniques de décomposition des attributs :** Ces techniques tentent d'expliquer la contribution de chaque caractéristique à une décision spécifique, en donnant une idée des caractéristiques les plus influentes.

Quoi qu'il en soit, les avantages de l'approche XAI sont clairs : elle renforce la **confiance** (lorsque les utilisateurs, en particulier ceux qui ne sont pas des experts en IA, comprennent le fonctionnement d'un système, ils sont plus susceptibles de lui faire confiance) ; elle contribue à maintenir la **responsabilité** (l'IA explicable peut contribuer à garantir que les systèmes d'IA agissent de manière responsable et équitable, en réduisant les biais et les erreurs) ; elle facilite l'**amélioration** et le **perfectionnement** des modèles (en comprenant la façon dont un modèle prend des décisions, il est plus facile d'identifier et de corriger les erreurs ou les imprécisions) ; et facilite la **conformité réglementaire** (puisque dans certaines juridictions, telles que l'UE, les systèmes de prise de décision automatisés doivent être transparents et capables de justifier leurs décisions).

Naturellement, l'application des modèles XAI n'est pas sans poser des problèmes, tels que le **choix entre précision et interprétabilité** (sachant qu'il existe souvent un équilibre entre la précision du modèle et son interprétabilité, les modèles plus simples peuvent être plus faciles à comprendre mais moins précis) ; la **subjectivité** (puisque l'"explicabilité" peut être subjective, c'est-à-dire que ce qui est clair et compréhensible pour un expert technique peut ne pas l'être pour un autre ou pour un non-spécialiste) ; ou la **généralisabilité** (puisque les explications relatives à une décision spécifique peuvent ne pas se généraliser à d'autres cas).

EXEMPLES :

LIME (Local Interpretable Model-agnostic Explanations) :
Il s'agit d'une technique permettant d'expliquer les prédictions de tout classificateur ou régresseur d'une manière compréhensible par l'homme. Elle consiste à créer un modèle interprétable qui est localement fidèle aux prédictions du modèle original. (https://github.com/marcotcr/lime)

SHAP (SHapley Additive exPlanations) :
Basé sur la théorie des jeux pour expliquer le résultat de n'importe quel modèle de machine. Il s'agit d'une mesure unifiée de l'importance des caractéristiques. (https://github.com/slundberg/shap)

DeepLIFT (Deep Learning Important FeaTures) :
Présente une approche permettant de décomposer les sorties des réseaux de neurones et de calculer l'importance de chaque entrée pour la sortie. Il est particulièrement utile pour les réseaux de neurones profonds. (https://github.com/kundajelab/deeplift)

AI Explainability 360 :
Il s'agit d'une boîte à outils comprenant des algorithmes, des bibliothèques et des tutoriels pour aider les développeurs à comprendre, expliquer et visualiser les décisions prises par les modèles d'IA. (https://aix360.mybluemix.net/)

InterpretML :
Bibliothèque open source de Microsoft pour l'interprétation des modèles de machines. Elle fournit une variété de techniques et d'outils pour l'interprétation des modèles. (https://interpret.ml/)

Adoption de la blockchain pour la sécurité

Bien que la blockchain soit surtout connue comme la technologie derrière le bitcoin, elle peut être utilisée pour la cybersécurité, en particulier dans la gestion d'identités et la sécurisation des données.

Voici quelques exemples de la manière dont la blockchain peut être utilisée pour la cybersécurité :

1. **Intégrité et authentification des données :** La blockchain fournit un registre immuable et transparent des données. Une fois qu'un bloc est ajouté à la chaîne, il ne peut être modifié sans modifier tous les blocs suivants (ce qui est extraordinairement difficile en raison de la nature décentralisée du réseau blockchain). Cela garantit l'intégrité des données et empêche toute falsification.
2. **Décentralisation :** La cybersécurité traditionnelle repose sur des serveurs centralisés et ceux-ci sont des points d'attaque vulnérables. La blockchain est intrinsèquement décentralisée, ce qui signifie qu'il n'y a pas un point unique de compromission potentielle. Ceci peut être un avantage ou une source de risque potentiel, tout dépend de la sécurité adoptée pour chaque nœud.
3. **Identité sécurisée :** Les systèmes basés sur la blockchain peuvent fournir des solutions d'identité numérique où les identités des utilisateurs sont vérifiées et stockées sur la blockchain.
4. **Communications sécurisées :** Les solutions blockchain peuvent garantir des communications sécurisées et authentifiées entre les appareils connectés à Internet (IoT). Ces appareils sont souvent vulnérables aux attaques, mais avec une gestion de l'identité basée sur la blockchain ils pourraient être validés et communiquer entre eux de manière plus transparente.
5. **Audit et traçabilité :** La blockchain fournit une trace claire et vérifiable de toutes les transactions. C'est un atout inestimable pour le processus d'audit car elle facilite la transparence et la responsabilité.
6. **Résistance à la censure et disponibilité :** En raison de leur nature décentralisée, les réseaux blockchain sont résistants à la censure et aux perturbations. Il est difficile de fermer ou de censurer un réseau blockchain sans le consensus de la majorité de ses participants.
7. **Contrats intelligents pour une automatisation sécurisée :** Les contrats intelligents sont des contrats numériques stockés dans une blockchain qui sont automatiquement exécutés lorsque des conditions générales prédéterminées sont remplies. Ils peuvent être utilisés pour automatiser l'exécution d'un accord sans intervention d'un intermédiaire, réduisant ainsi l'exposition au risque de fraude ou la possibilité d'une intervention malveillante.

Bien que la blockchain soit surtout connue comme la technologie derrière le bitcoin, elle peut être utilisée pour la cybersécurité, en particulier dans la gestion d'identités et la sécurisation des données

Malgré les avantages de la technologie blockchain pour la cybersécurité, il y a aussi des défis à relever. Par exemple, bien que la blockchain soit immuable et que les transactions ne puissent pas être modifiées une fois validées, si un attaquant parvient à prendre le contrôle de la majorité du réseau (attaque à 51 %), il pourrait potentiellement valider des transactions frauduleuses. En outre, comme toute technologie émergente, la mise en œuvre pratique de la blockchain dans le domaine de la cybersécurité est encore en cours de développement, et il est donc conseillé que les organisations fassent preuve de prudence et de diligence en l'adoptant.

EXEMPLES :

Guardtime :
Utilise la technologie blockchain pour garantir l'intégrité et l'authenticité des données.
(https://www.guardtime.com/)

Civic :
Il s'agit d'une solution d'identité sécurisée basée sur la blockchain. Elle fournit aux entreprises et aux particuliers des outils pour contrôler et protéger les identités. Grâce à sa plateforme décentralisée, Civic permet une authentification sans avoir recours aux mots de passe traditionnels.
(https://www.civic.com/)

REMME :
Il s'agit d'une solution qui vise à éliminer les attaques phishing, les mots de passe et les certificats. Elle utilise la blockchain —au lieu d'un mot de passe— pour authentifier les utilisateurs et les appareils.
(https://remme.io/)

Chaîne des objets (CoT) :
Recherche et développe des cas d'usage combinant la blockchain et l'Internet des objets (IoT), ce qui peut être avantageux dans des domaines tels que l'énergie, le transport et la logistique, où l'intégrité et la sécurité des données collectées par les IoT sont essentielles.
(https://www.chainofthings.com/)

Modèles d'IA basés sur le comportement de l'utilisateur

Au lieu de s'appuyer uniquement sur des mots de passe ou des données biométriques, l'IA pourrait analyser des modèles de comportement habituel (par exemple la vitesse à laquelle une personne tape ou la façon dont elle déplace sa souris) pour authentifier les utilisateurs et détecter des comportements anormaux.

Les modèles d'IA basés sur le comportement des utilisateurs constituent une tendance émergente en matière de cybersécurité et représentent l'un des moyens les plus avancés de détecter les anomalies et les activités suspectes dans un système. Comme nous l'avons dit, ces modèles sont entrainés pour apprendre les modèles normaux de comportement de l'utilisateur afin d'identifier tout écart par rapport à ces modèles comme une activité potentiellement suspecte.

Les **cas d'usage** les plus communs de ce type de modèle sont les suivants :

1. **Détection des fraudes :** Dans le secteur financier, par exemple, si un utilisateur effectue des transactions à montant élevé de façon soudaine ou si les schémas ne correspondent pas à son comportement habituel, l'IA peut générer une alerte.
2. **Contrôle d'accès et authentification :** Si le comportement d'un utilisateur change soudainement (par exemple, s'il se connecte à des heures inhabituelles ou à partir de lieux géographiques inconnus), cela peut être le signe que quelqu'un d'autre utilise ses identifiants.
3. **Protection contre les menaces internes :** Les employés mécontents ou malveillants peuvent constituer une menace pour les organisations. S'ils commencent à accéder à des fichiers ou à des systèmes qu'ils n'utilisent pas normalement, les systèmes d'IA basés sur l'analyse comportementale peuvent le détecter.

L'utilisation de ces modèles présente des **avantages** indéniables, notamment : la **personnalisation** (puisqu'ils sont basés sur le comportement individuel de l'utilisateur, ils sont hautement personnalisés et adaptatifs) ; la **détection proactive** (ils peuvent détecter les menaces en temps réel, ce qui permet une réaction plus rapide) ; et la **réduction des faux positifs** (le volume d'alertes se réduit puisque les règles de détection sont ajustées au comportement réel de l'utilisateur et elles ne signalent pas à tort les activités bénignes qui s'écartent légèrement de la norme).

Toutefois, l'utilisation de ces modèles pose également certains **défis**, tels que : le **temps d'apprentissage** (comme tout système d'apprentissage automatique, ces modèles ont besoin de temps pour apprendre les modèles de comportement de l'utilisateur) ; les **changements de comportement de l'utilisateur** (si un utilisateur change de rôle ou de responsabilité, son comportement peut changer, ce qui peut entraîner des faux positifs jusqu'à ce que le système s'adapte) ou la **protection de la vie privée** (puisque ces systèmes collectent et analysent de nombreuses informations sur le comportement de l'utilisateur, ce qui soulève des questions sur la protection de la vie privée et sur la manière dont ces données sont traitées et stockées).

L'IA quantique

À mesure que l'informatique quantique deviendra une réalité plus quotidienne, nous assisterons probablement à des développements spécifiques dans le domaine de l'IA quantique. Ces systèmes pourraient être capables de traiter des données à des vitesses exponentielles et de traiter des problèmes de sécurité qui sont actuellement insolubles pour les systèmes conventionnels. L'"IA quantique" est un domaine émergent qui combine des techniques et des concepts de l'intelligence artificielle avec la mécanique quantique et l'informatique quantique. Bien qu'il en soit encore à ses débuts, ce domaine promet de révolutionner les capacités et les performances des systèmes d'intelligence artificielle.

Comme chacun sait, l'informatique quantique est basée sur la mécanique quantique, une théorie de la physique qui décrit le fonctionnement des particules subatomiques. Contrairement à l'informatique classique, qui utilise des bits (0 et 1) pour représenter et traiter l'information, l'informatique quantique utilise des "qubits". Les qubits ont la capacité étonnante de représenter plusieurs états à la fois (superposition) et d'être "intriqués", ce qui signifie que l'état d'un qubit peut dépendre de l'état d'un autre, quelle que soit la distance qui les sépare.

EXEMPLES :

Darktrace :
Dont nous avons déjà parlé, utilise ce qu'il appelle un "système immunitaire d'entreprise" pour apprendre à partir du "mode de vie" normal de chaque utilisateur et de chaque appareil sur un réseau, puis pour identifier les comportements anormaux. (https://www.darktrace.com)

Cylance :
Comme indiqué ci-dessus, cette société utilise l'IA pour offrir une solution de prévention des menaces pour les points finaux basée sur le comportement. Sa plateforme a pour but de stopper les malware, les ransomware et d'autres menaces en se basant sur l'identification de comportements suspects plutôt que sur des signatures de virus connues. (https://www.cylance.com)

Exabeam :
Il s'agit d'une plateforme de gestion des informations et des événements de sécurité (SIEM) qui utilise l'apprentissage automatique pour analyser le comportement des utilisateurs et détecter les menaces. (https://www.exabeam.com)

UserInsight de Rapid7 :
Il se concentre spécifiquement sur la détection des comportements anormaux des utilisateurs et des attaquants au sein d'un réseau. Il peut identifier les cas où les identifiants d'un utilisateur ont été compromis et sont utilisés par un attaquant. (https://www.rapid7.com)

Ces outils combinent des techniques de sécurité traditionnelles avec un apprentissage automatique avancé pour détecter les menaces en temps réel sur la base du comportement de l'utilisateur.

L'utilisation de modèles d'IA quantique présenterait les **avantages** suivants :

- **Vitesse et évolutivité :** les algorithmes quantiques peuvent effectuer certaines opérations beaucoup plus rapidement que leurs homologues classiques. En théorie, l'IA quantique pourrait s'attaquer à des problèmes qui sont actuellement insolubles pour les ordinateurs classiques en raison de leur complexité.
- **Optimisation :** des problèmes tels que l'optimisation combinatoire, qui sont cruciaux dans des domaines tels que la logistique, l'économie et de nombreuses applications de l'IA, pourraient grandement bénéficier de la capacité des ordinateurs quantiques à explorer plusieurs solutions simultanément.
- **Apprentissage en profondeur et formation de modèles :** les ordinateurs quantiques ont le potentiel d'accélérer considérablement la formation de modèles d'IA complexes, ce qui pourrait révolutionner des domaines tels que l'apprentissage profond.

Cependant, comme toujours, l'utilisation de ces modèles pose certains **défis :**

- **Matériel :**
  Bien que des progrès significatifs aient été réalisés dans la construction d'ordinateurs quantiques, nous sommes toujours confrontés à des défis en termes de stabilité, de cohérence et d'évolutivité.
- **Algorithmes quantiques :**
  L'adaptation des algorithmes d'IA classiques à l'informatique quantique reste un domaine de recherche actif. Tous les problèmes ne bénéficieront pas d'une solution quantique.
- **Interaction quantique-classique :**
  L'intégration de systèmes informatiques quantiques avec des infrastructures et des algorithmes classiques est un défi considérable.

EXEMPLES :

IBM Q Experience :
IBM est un leader dans le domaine de l'informatique quantique et, par l'intermédiaire d'IBM Q, offre aux chercheurs et aux développeurs l'accès à de véritables ordinateurs quantiques pour expérimenter et développer des algorithmes quantiques. Bien qu'elle ne soit pas exclusivement destinée à l'IA, cette plateforme pourrait être utilisée pour expérimenter des algorithmes d'IA quantique.
(https://quantum-computing.ibm.com/)

D-Wave Systems :
D-Wave est connue pour ses systèmes quantiques avancés, différents des modèles informatiques basés sur les portes quantiques. Elle a travaillé sur l'optimisation et l'apprentissage automatique à l'aide de ses systèmes quantiques.
(https://www.dwavesys.com/).

Google AI Quantum :
Google a mené des recherches actives dans le domaine de l'informatique quantique et a fait des progrès significatifs. Bien qu'il se concentre sur de nombreux aspects de l'informatique quantique, l'intelligence artificielle est
l'une des applications qu'il explore.
(https://quantumai.google/)

Rigetti Computing :
Il s'agit d'une start-up qui se concentre sur la construction d'ordinateurs quantiques et fournit également une plateforme en nuage pour les développeurs et les scientifiques afin d'expérimenter l'informatique quantique. Bien que la plateforme ne soit pas exclusivement dédiée à l'IA quantique, elle offre la possibilité d'explorer des applications dans ce domaine.
(https://www.rigetti.com/)

Il est important de noter qu'un grand nombre de ces outils et systèmes sont conçus comme des plateformes pour l'informatique quantique en général, et non spécifiquement pour l'IA quantique. Cependant, étant donné que l'informatique quantique a des applications potentielles dans l'IA, ces plateformes peuvent jouer un rôle crucial dans le développement futur de l'IA quantique.

La collaboration Homme-Machine

Malgré les progrès de l'IA, les humains resteront essentiels dans le domaine de la cybersécurité. La nouvelle tendance consistera à mettre en place des systèmes dans lesquels les humains et les machines travailleront ensemble, en complétant leurs forces respectives.

En effet, la collaboration homme-machine est une approche qui exploite les capacités uniques des humains et des machines, en particulier dans le contexte de l'intelligence artificielle, afin d'améliorer la prise de décision, l'efficacité et les résultats globaux. L'idée principale qui sous-tend cette collaboration est que si les machines sont excellentes pour calculer, analyser et traiter de grands volumes de données, les humains possèdent l'intuition, la compréhension du contexte, l'empathie et la créativité.

Cette symbiose entre l'homme et la machine comporte certains **aspects essentiels** qu'il convient de garder à l'esprit :

1. **Complémentarité :** L'IA et l'homme ont des caractéristiques complémentaires. Par exemple, l'IA peut traiter de grandes quantités de données, effectuer des opérations répétitives et des calculs complexes à des vitesses que les humains ne peuvent pas atteindre. Cependant, les humains apportent leur créativité, leur jugement et leur expérience basée sur l'intuition.
2. **Interaction naturelle :** Le développement d'interfaces intuitives et naturelles, telles que la conversation vocale ou l'interaction gestuelle, permet une collaboration plus transparente entre les machines et les humains. Ces systèmes s'appuient sur le traitement du langage naturel, la reconnaissance vocale et la reconnaissance gestuelle pour comprendre les interactions humaines et y répondre de manière appropriée.
3. **Apprentissage à double sens :** Tandis que la machine apprend du comportement humain pour améliorer ses prévisions et ses actions, l'homme apprend également à faire confiance à la machine et à comprendre son fonctionnement, établissant ainsi un cycle de retour d'information et d'amélioration continue.
4. **Transparence et explicabilité :** Pour que les humains fassent confiance aux décisions prises ou suggérées par les machines, il est essentiel que ces dernières puissent fournir des explications compréhensibles sur leurs décisions.

La nouvelle tendance consistera à mettre en place des systèmes dans lesquels les humains et les machines travailleront ensemble, en complétant leurs forces respectives

5. **Intervention humaine :** Dans de nombreux systèmes collaboratifs, un mécanisme est intégré au système pour permettre l'intervention humaine dans certains scénarios. Par exemple, dans un système de conduite autonome, il peut y avoir des situations où le système demande au conducteur humain de prendre le contrôle.

EXEMPLES :

IBM Watson :
L'une des utilisations les plus remarquables de Watson est le domaine médical. *Watson Health* aide les professionnels de la santé à prendre des décisions éclairées sur le traitement des patients en analysant de grandes quantités de données médicales et de littérature scientifique. Bien que Watson fournisse des recommandations, c'est toujours le médecin qui prend la décision finale. (https://www.merative.com/)

Google's DeepMind AlphaGo :
Bien qu'il soit surtout connu pour avoir battu des champions humains au jeu de Go, le véritable exploit réside dans la manière dont les humains et les machines ont appris les uns des autres, ce qui a permis aux joueurs de Go d'étudier les mouvements d'AlphaGo pour améliorer leurs propres stratégies. (https://www.deepmind.com/research/highlighted-research/alphago)

KUKA LBR iiwa :
Robot collaboratif conçu pour travailler aux côtés de l'homme dans un environnement industriel. Ces "cobots" sont sensibles au toucher et peuvent s'arrêter ou ralentir s'ils détectent un objet ou une personne sur leur chemin, ce qui permet aux humains et aux robots de travailler côte à côte sur la même tâche. (https://www.kuka.com/en-us/products/robotics-systems/industrial-robots/lbr-iiwa)

OpenAI Codex :
Plateforme basée sur l'IA qui aide les programmeurs à écrire du code. Elle peut générer des extraits de code à partir de descriptions fournies par l'utilisateur, agissant comme un assistant de programmation. (https://openai.com/blog/openai-codex)

Adobe Sensei :
Intégré aux outils Adobe, Sensei utilise l'IA et l'apprentissage automatique pour faciliter les tâches créatives, de l'édition de photos et de vidéos à la conception et à l'illustration. Bien qu'il automatise certaines fonctions, le créatif garde le contrôle final et utilise l'outil pour améliorer et accélérer le processus de conception. (https://www.adobe.com/sensei.html)

Ces exemples représentent une variété d'applications dans différents secteurs et ils nous montrent comment les outils basés sur l'IA peuvent travailler en collaboration avec les humains pour améliorer l'efficacité, la précision et la créativité.

L'IA en périphérie du réseau (Edge AI)

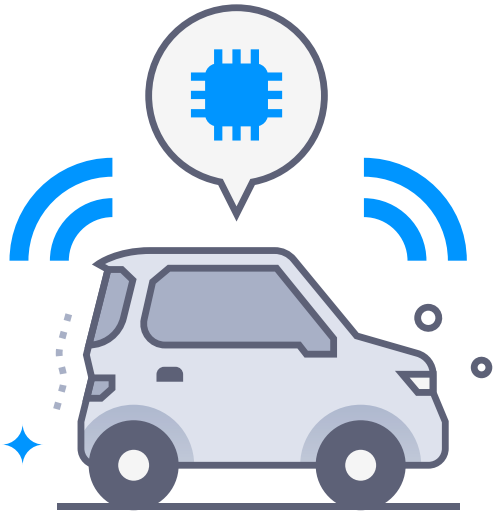
Au lieu de s'appuyer sur des centres de données centralisés, l'IA pourrait être traitée sur l'appareil (téléphones mobiles, IoT, etc.). Cela a des implications significatives pour la cybersécurité, permettant des réponses plus rapides et réduisant les risques associés à la transmission des données.

L'utilisation de ce type de modèle présente les **avantages** suivants :

1. **Réduction de la latence :** Le traitement des données en local, au lieu de les envoyer dans le cloud, permet de réduire le temps de latence. Cet aspect est particulièrement important car l'Edge AI se prête très bien aux usages en temps réel, à commencer par les voitures autonomes.
2. **Confidentialité et sécurité :** Le fait de traiter les données localement sur un appareil permet de minimiser les risques de sécurité liés à la transmission des données et de garantir que les données sensibles ne quittent pas l'appareil.
3. **Fonctionnement hors ligne :** Les appareils dotés de capacités Edge AI sont capables de fonctionner et de prendre des décisions sans besoin d'une connexion active au Cloud ou au serveur central.
4. **Efficacité de la bande passante :** Diminution de l'utilisation de la bande passante dans le flux de données grâce au traitement, à l'analyse et au stockage des données localement au lieu de les envoyer dans le cloud.
5. **Consommation d'énergie :** Si les appareils d'Edge AI peuvent nécessiter plus de puissance énergétique que les dispositifs dépourvus de capacités d'IA, ils consomment souvent moins d'énergie que ce qui est nécessaire pour communiquer en permanence avec un serveur central.

Au lieu de s'appuyer sur des centres de données centralisés, l'IA pourrait être traitée sur l'appareil (téléphones mobiles, IoT, etc.)

L'Edge AI se prête par définition très bien aux usages IoT, mais également dans l'ensemble des appareils mobiles : les **voitures autonomes** (étant donné que les voitures doivent traiter rapidement d'énormes quantités de données provenant de leurs capteurs pour naviguer en toute sécurité, la latence dans la prise de décision peut avoir de graves conséquences) ; les **appareils ménagers intelligents** (réfrigérateurs, aspirateurs, fours et autres appareils qui utilisent l'IA pour l'optimisation et la prise de décision) ; les **dispositifs médicaux portables** (moniteurs de fréquence cardiaque, appareils de mesure du glucose et autres dispositifs médicaux qui doivent traiter des données en temps réel) ; les **caméras de sécurité** (qui détectent les activités anormales ou reconnaissent les visages et prennent des décisions sur la base de ces données) ou les drones (utilisés pour la navigation, la détection d'objets et la prise de décisions en temps réel).



De l'autre côté de l'échelle, les **défis** liés à ce type de modèle sont les suivants :

- **Limitations matérielles :** Bien que les dispositifs Edge évoluent rapidement, il existe encore des limitations en termes de capacité de traitement, de mémoire et de stockage par rapport aux centres de données centralisés.
- **Gestion et mise à jour :** La maintenance et la mise à jour des modèles d'IA sur de nombreux appareils dispersés peuvent constituer un défi.
- **Développement de modèles :** Les modèles doivent souvent être optimisés et adaptés pour être suffisamment petits et efficaces pour fonctionner sur des appareils Edge sans sacrifier indûment la précision ou la capacité.

Toutes ces tendances émergentes en matière d'IA pour la cybersécurité ne représentent qu'une petite partie de ce qui est sans doute à venir.

EXEMPLES :

TensorFlow Lite :
Une solution Google conçue pour apporter des modèles d'apprentissage automatique aux appareils mobiles et Edge. Elle fournit des outils pour convertir et optimiser les modèles TensorFlow standard afin qu'ils soient efficaces sur ces appareils. (https://www.tensorflow.org/lite)

ONNX Runtime :
Ce moteur comprend une bibliothèque d'inférence d'apprentissage automatique pour les modèles ONNX (Open Neural Network Exchange). ONNX Runtime est léger et peut être utilisé sur différentes plateformes, y compris les appareils Edge. (https://onnxruntime.ai/)

Intel OpenVINO (Open Visual Inferencing & Neural Network Optimization) Toolkit :
Le toolkit d'Intel pour l'accélération et l'optimisation des modèles d'intelligence artificielle, spécialement conçu pour fonctionner efficacement sur le hardware Intel, y compris les puces conçues pour les appareils Edge. (https://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit.html)

NVIDIA Jetson Platform :
Une série de systèmes embarqués conçus par NVIDIA qui intègrent un GPU et sont optimisés pour les tâches d'inférence AI Edge. (https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/)

Azure IoT Edge :
Une solution Microsoft qui permet de déployer des charges de travail basées sur le cloud directement sur l'Internet des objets (IoT) et d'autres appareils périphériques, y compris des modèles d'apprentissage automatique. (https://azure.microsoft.com/en-us/services/iot-edge/)

Ces outils et plateformes ne représentent qu'une fraction de l'industrie croissante de l'IA Edge.

6.2 Enquêtes en cours

Le domaine de la cybersécurité est un terrain fertile pour la recherche, en particulier avec l'intégration de technologies émergentes telles que l'intelligence artificielle. Cette section explore les domaines de recherche les plus récents.

apprentissage profond contre les attaques réseau avancées :	Les attaques réseau évoluent pour devenir plus sophistiquées et plus difficiles à détecter. Les recherches actuelles portent sur la manière d'utiliser les techniques d'apprentissage profond, telles que les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents (RNN), pour identifier des schémas camouflés dans le trafic réseau pouvant indiquer une attaque.
Techniques adverses pour renforcer les systèmes d'intelligence artificielle :	L'idée est d'utiliser des techniques adverses (attaques spécifiquement conçues pour tromper les modèles d'IA) dans un environnement contrôlé afin d'améliorer la robustesse des modèles d'IA dans le domaine de la cybersécurité. Il s'agit d'entraîner les modèles avec des données perturbées pour les rendre plus résistants aux attaques adverses dans des scénarios réels.
Sécurité de la chaîne d'approvisionnement en IA :	L'IA devenant un élément essentiel de nombreux systèmes, il est devenu crucial de garantir l'intégrité de l'ensemble de la chaîne d'approvisionnement de l'IA, de l'entraînement au déploiement. La recherche se concentre sur la manière dont ces systèmes peuvent être infiltrés et compromis et sur la manière de se défendre contre ces vulnérabilités.
Détection des "deepfakes" et, en général, des contenus synthétiques :	La capacité croissante des outils d'IA à créer des contenus synthétiques réalistes (comme les deepfakes audio ou vidéo, par exemple) a entraîné une intensification de la recherche sur la détection automatique de ces contenus. Cela a des applications directes en matière de cybersécurité, en particulier dans des domaines tels que l'authentification et la protection contre la désinformation.
Automatisation de la réponse aux incidents :	Plutôt que de simplement détecter les menaces, des systèmes d'IA sont développés pour répondre automatiquement aux incidents de sécurité, en prenant des décisions en temps réel sur la façon d'atténuer ou de neutraliser une menace.

Comprendre la "boîte noire" de l'IA :	L'IA explicable (XAI) est un domaine de recherche majeur. Dans le contexte de la cybersécurité, il est essentiel de comprendre pourquoi un système d'IA prend une décision particulière, surtout si elle est liée à la détection d'une menace ou à la réponse à un incident.
IA pour la protection contre les menaces internes :	Les menaces internes, qu'elles soient malveillantes ou inaperçues, restent un défi majeur en matière de cybersécurité. L'IA peut jouer un rôle dans la surveillance du comportement des utilisateurs afin d'identifier les activités suspectes ou les écarts par rapport à la norme.
Méthodes d'entraînement sûres:	Étudier les moyens d'entraîner les modèles d'IA sans exposer les données sensibles, comme l'apprentissage fédéré ou la confidentialité différentielle.

Ces domaines de recherche témoignent de l'évolution et de l'adaptation continues de l'IA dans le domaine de la cybersécurité. Les menaces évoluant et devenant plus sophistiquées, il est essentiel que la recherche dans ce domaine suive le rythme pour fournir des solutions efficaces et proactives.

EXEMPLES :

Apprentissage profond contre les attaques réseau avancées :
"DeepDefense", une plateforme qui utilise des techniques d'apprentissage profond pour détecter les attaques en temps réel.
(https://ieeexplore.ieee.org/document/7946998)

Techniques adverses pour renforcer les systèmes d'IA :
Le projet "CleverHans" de Google.
(GitHub.com/tensorflow/cleverhans)

Sécurité de la chaîne d'approvisionnement en IA :
La MITRE Corporation a mené des recherches dans des domaines connexes.
(mitre.org).

Détection des deepfakes et des contenus synthétiques :
"Deepware Scanner" de la société Cyabra.
(cyabra.com)

Automatisation de la réponse aux incidents :
"Cortex XSOAR" de Palo Alto Networks.
(paloaltonetworks.com/cortex/xsoar)

Comprendre la "boîte noire" de l'IA :
LIME (Local Interpretable Model-Agnostic Explanations).
(github.com/marcotcr/lime)

IA pour la protection face aux menaces internes :
"Insider Threat Solution" de Varonis.
(varonis.com/solutions/insider-threat-detection)

Méthodologies d'entraînement sûres :
"TensorFlow Federated" pour l'apprentissage fédéré.
(tensorflow.org/federated)

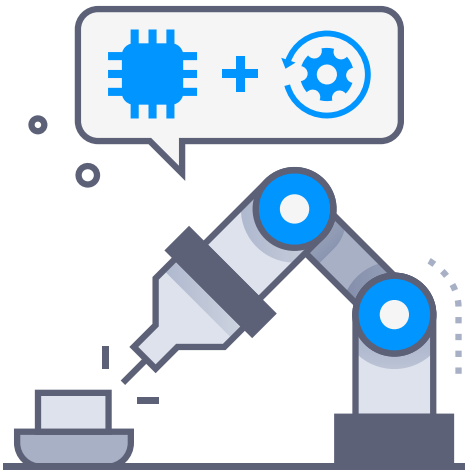
6.3 Impact potentiel sur l'industrie et la société

Les progrès de l'IA dans le domaine de la cybersécurité pourraient redéfinir de nombreux aspects de l'industrie et exercer une influence notable sur la société. À mesure que les systèmes autonomes deviennent plus sophistiqués, l'**impact** se fera sentir à différents niveaux :

1. **Optimiser la sécurité des entreprises :**
   - Les entreprises peuvent s'attendre à une meilleure protection contre les menaces grâce à des systèmes capables d'apprendre et de s'adapter en temps réel.
   - On s'attend à une réduction des délais de réponse aux incidents et à une amélioration de la capacité à prévenir les failles de sécurité avant qu'elles ne se produisent.
2. **Transformer le travail des professionnels de la cybersécurité :**
   - L'IA peut prendre en charge les tâches de routine, ce qui permet aux professionnels de la cybersécurité de se concentrer sur des tâches plus stratégiques ou plus complexes.
   - Cela pourrait entraîner une restructuration des rôles et des responsabilités, ainsi que la nécessité d'acquérir de nouvelles compétences et de nouvelles formations.
3. **Une société numériquement plus sûre :**
   - À mesure que les solutions basées sur l'IA sont intégrées dans un plus grand nombre de plateformes et de services, le citoyen moyen peut bénéficier d'une sécurité numérique plus solide dans sa vie quotidienne.
   - Les transactions en ligne, le stockage de données personnelles et d'autres activités numériques peuvent présenter moins de risques.

Les progrès de l'IA dans le domaine de la cybersécurité pourraient redéfinir de nombreux aspects de l'industrie et exercer une influence notable sur la société

4. **Changements dans la nature des menaces :**
   - Les attaquants évolueront également et adapteront leurs méthodes en réponse à des systèmes de défense plus avancés.
   - Nous pourrions assister à une augmentation des attaques hautement sophistiquées et ciblées qui utilisent l'IA pour trouver et exploiter les vulnérabilités.
5. **Défis éthiques et protection de la vie privée :**
   - À mesure que l'IA devient un outil courant en matière de cybersécurité, des inquiétudes se font jour quant à l'utilisation et à l'abus des données personnelles.
   - Les entreprises et les organisations doivent être transparentes sur la manière dont elles utilisent l'IA et veiller à ce que les droits de confidentialité soient respectés.
6. **Économie et marché du travail :**
   - L'adoption massive de solutions d'IA pourrait influencer la demande de professionnels de la cybersécurité, en augmentant éventuellement la demande d'experts spécialisés dans l'IA et en diminuant le besoin de rôles plus traditionnels.
   - Les startups et les entreprises qui développent des solutions d'IA pour la cybersécurité pourraient connaître une croissance importante, influençant l'économie et créant de nouvelles opportunités de marché.
7. **Règles et règlements :**
   - Les gouvernements et les organismes régulateurs du monde entier —comme c'est actuellement le cas dans l'Union Européenne— pourraient introduire de nouvelles réglementations pour garantir que l'IA soit utilisée de manière responsable, notamment en matière de cybersécurité.
   - Ces réglementations pourraient influencer la manière dont les entreprises développent, mettent en œuvre et utilisent les solutions d'IA.



L'impact de l'IA sur la cybersécurité est vaste et couvre de nombreuses facettes de l'industrie et de la société. Il est essentiel que les parties prenantes, des professionnels aux régulateurs, soient informées et préparées à aborder ces changements de manière proactive et éthique.

7. Recommandations et bonnes pratiques

L'intelligence artificielle (IA) devenant de plus en plus pertinente dans le monde de la cybersécurité, tant pour la défense que pour l'attaque, il devient impératif que les organisations adoptent des stratégies éclairées pour sa mise en œuvre. Toutefois, l'IA n'est pas une solution miracle que l'on peut appliquer sans cérémonie ; son utilisation efficace nécessite une compréhension nuancée et une approche stratégique.

Dans cette section, nous allons explorer les recommandations et les bonnes pratiques que les organisations devraient prendre en compte lorsqu'elles intègrent des solutions d'IA dans leurs systèmes de cybersécurité. De la sélection et de l'entraînement des modèles à la mise en œuvre et à la surveillance de ceux-ci en temps réel, il est essentiel que les organisations soient équipées des meilleures pratiques pour s'assurer que l'IA devienne un atout et non une faiblesse.

Nous apprendrons comment garantir la robustesse et la fiabilité des systèmes d'IA, comment gérer et protéger les données qui nourrissent ces systèmes, et comment faire en sorte que l'éthique et la transparence soient au cœur de toute mise en œuvre de l'IA. Nous soulignerons également l'importance de la formation continue et de l'adaptabilité, compte tenu de la nature changeante des cybermenaces.

L'intelligence artificielle (IA) devenant de plus en plus pertinente dans le monde de la cybersécurité, tant pour la défense que pour l'attaque, il devient impératif que les organisations adoptent des stratégies éclairées pour sa mise en œuvre

7.1 Intégration des équipes de cybersécurité et des équipes d'IA

La convergence efficace de l'intelligence artificielle et de la cybersécurité nécessite non seulement la combinaison de technologies, mais aussi la collaboration entre des experts des deux domaines. L'intégration efficace de ces équipes peut renforcer les capacités de cyberdéfense d'une organisation et garantir que les solutions d'IA sont robustes, fiables et appropriées pour faire face aux cybermenaces réelles.

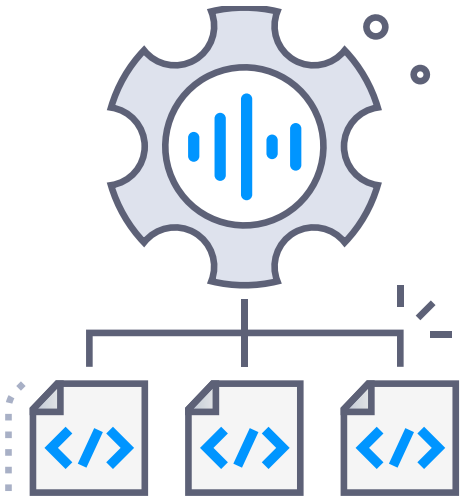
Les éléments essentiels à la réussite :

1. **Communication et collaboration :** Un flux de communication constant et efficace entre les équipes de cybersécurité et les équipes d'IA est essentiel. Les défis, les objectifs et les solutions des deux domaines doivent être partagés et compris des deux côtés.
2. **Formation croisée :** Le fait de dispenser une formation sur les principes fondamentaux de la cybersécurité à l'équipe chargée de l'IA et, réciproquement, sur les principes fondamentaux de l'IA à l'équipe chargée de la cybersécurité peut favoriser la compréhension mutuelle et améliorer la collaboration.
3. **Développement conjoint de solutions :** Plutôt que de travailler en silo, les équipes chargées de l'IA et de la cybersécurité devraient collaborer à la conception, au développement et au déploiement des solutions. Cela permet de s'assurer que les solutions d'IA sont pertinentes et alignées sur les objectifs de cybersécurité.
4. **Examens réguliers et retour d'information :** Les solutions d'IA dans le domaine de la cybersécurité doivent faire l'objet d'examens réguliers par les deux équipes. Ces examens permettent d'identifier les lacunes, les domaines à améliorer et les adaptations nécessaires pour faire face aux nouvelles menaces.
5. **Tests conjoints :** Comme pour le développement, la vérification (ou test) des solutions devrait également être une activité conjointe. Cela permettrait d'identifier et de corriger les bogues avant qu'ils ne soient exploités par des adversaires.

La convergence efficace de l'intelligence artificielle et de la cybersécurité nécessite non seulement la combinaison de technologies, mais aussi la collaboration entre des experts des deux domaines

6. **Intégration des outils et des plateformes :** L'utilisation d'outils et de plateformes permettant l'intégration et la collaboration entre les deux équipes peut s'avérer essentielle, notamment l'utilisation de plateformes de développement collaboratif, de systèmes de gestion de projet et d'outils de communication.
7. **Respect et appréciation mutuels :** Pour une collaboration efficace, il est essentiel que les deux équipes reconnaissent et apprécient leurs compétences et contributions respectives. L'intelligence artificielle et la cybersécurité sont des disciplines complexes et spécialisées, et il est essentiel qu'elles se respectent mutuellement pour garantir une collaboration efficace.
8. **Planification de scénarios de crise :** En cas d'incident de sécurité, il est essentiel de mettre en place des plans sur la manière dont les équipes travailleront ensemble. Il s'agit notamment de déterminer les rôles, les responsabilités et les flux de communication.
9. **Mises à jour et formation continue :** Les domaines de l'IA et de la cybersécurité étant en constante évolution, il est essentiel que les deux équipes se tiennent au courant des dernières tendances, techniques et menaces dans leurs domaines respectifs.

En fin de compte, l'intégration efficace des équipes de cybersécurité et d'IA n'est pas simplement une option, mais une nécessité dans le monde d'aujourd'hui. Les cybermenaces évoluent rapidement et la combinaison de l'expertise en cybersécurité avec des capacités d'IA avancées peut constituer une défense solide contre des adversaires de plus en plus sophistiqués. Toutefois, pour que cette collaboration soit fructueuse, il est essentiel d'adopter de bonnes pratiques pour favoriser la communication, la compréhension et la collaboration entre ces équipes spécialisées.



7.2 Formation continue

Dans un monde où la technologie et les cybermenaces évoluent à un rythme vertigineux, la formation continue est essentielle pour rester à jour et assurer une défense efficace contre les adversaires. Cet impératif s'applique non seulement aux professionnels de la cybersécurité, mais aussi à ceux qui travaillent à l'intersection de l'IA et de la cybersécurité.

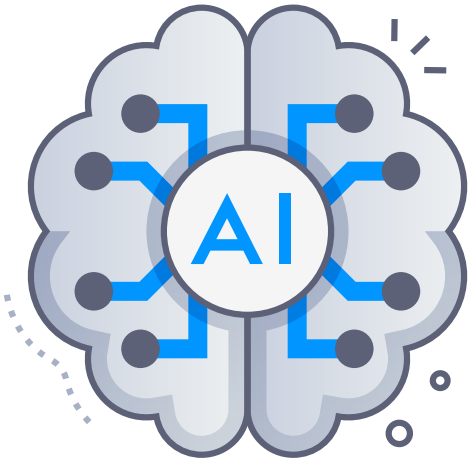
Pour y parvenir, il est conseillé de prendre en compte :

1. **Programmes de formation actualisés :** Les établissements d'enseignement et les organismes de formation devraient régulièrement revoir et actualiser leurs programmes afin de tenir compte des dernières évolutions en matière de cybersécurité et d'IA. Les nouveaux professionnels disposeront ainsi des connaissances les plus récentes.
2. **Ateliers et séminaires :** L'organisation d'ateliers et de séminaires sur des sujets émergents, ou la participation à ces ateliers et séminaires, peut permettre d'acquérir une connaissance approfondie de techniques spécifiques, de menaces émergentes ou de nouvelles approches dans le domaine de la cybersécurité et de l'IA.
3. **Certifications professionnelles :** Les certifications telles que CISSP, CISM, et d'autres liées à l'IA et/ou à la cybersécurité, peuvent aider les professionnels à valider et à améliorer leurs compétences. Ces certifications exigent souvent une formation continue, ce qui permet aux professionnels de rester à jour.
4. **Formation en cours d'emploi :** Les organisations devraient favoriser une culture de l'apprentissage continu, en offrant aux employés la possibilité d'être formés aux nouveaux outils, techniques et meilleures pratiques.
5. **Participation à des communautés et à des forums :** L'adhésion active à des communautés en ligne et à des forums spécialisés peut constituer une plateforme permettant d'apprendre de collègues, de partager des connaissances et de se tenir au courant des derniers développements et défis.

Dans un monde où la technologie et les cybermenaces évoluent à un rythme vertigineux, la formation continue est essentielle pour rester à jour et assurer une défense efficace contre les adversaires

6. **Simulations et exercices pratiques :** La réalisation de simulations et d'exercices pratiques, tels que la "capture the flag" ou les wargames, peut aider les professionnels à appliquer leurs connaissances à des scénarios réels, à améliorer leurs compétences et à apprendre de leurs erreurs.
7. **Ressources en ligne et MOOC :** Avec la prolifération des plateformes d'enseignement en ligne, il existe une abondance de cours (souvent gratuits) couvrant une grande variété de sujets dans les domaines de l'IA et de la cybersécurité. Ces cours peuvent être un excellent moyen d'apprendre à son propre rythme.
8. **Participation à des conférences :** Les conférences telles que Black-Hat, DEF-CON, STIC CCN-CERT et d'autres conférences spécifiques à l'IA offrent la possibilité de s'informer sur les recherches, les découvertes et les développements récents dans ce domaine.

En bref, la formation continue en matière de cybersécurité et d'IA ne sont pas un luxe, mais une nécessité. Pour mettre en place une défense efficace contre les menaces actuelles et futures, les professionnels doivent constamment mettre à jour et développer leurs connaissances et leurs compétences. Les organisations et les professionnels qui investissent dans la formation continue seront mieux placés pour faire face aux risques et les atténuer dans le paysage évolutif de la cybersécurité.



Vous trouverez ci-dessous une liste de références en matière de formation couvrant à la fois l'intelligence artificielle (IA) et la cybersécurité. Il s'agit d'une petite liste de références comprenant des cours, des certifications et des ressources.

1. **Cours et spécialisations :**
   - **Coursera:**
     - Spécialisation en cybersécurité
     - Introduction à l'intelligence artificielle
     - Spécialisation en apprentissage profond
   - **edX:**
     - Fondements de la cybersécurité
     - Principes de l'intelligence artificielle
   - **Udacity:**
     - Nanodegree en cybersécurité
     - Nanodegree en intelligence artificielle

2. **Certifications :**
   - **Certification professionnelle en cybersécurité (CISSP) :**
     Une certification mondialement reconnue qui démontre les capacités et les connaissances en matière de cybersécurité. Plus d'informations sur le site officiel de **l'ISC^2**.
   - **Certified Information Security Manager (CISM) :**
     Proposée par l'ISACA, cette certification est essentielle pour la gestion de la sécurité de l'information. Plus de détails sur le site officiel de **l'ISACA**.
   - **TensorFlow Developer Certificate :**
     Une certification axée sur l'IA et l'apprentissage profond. Plus d'informations sur le site officiel de **TensorFlow**.

3. **Ressources complémentaires :**
   - **MIT OpenCourseWare :**
     - Introduction à l'apprentissage profond
     - Informatique et sécurité des réseaux
   - **Cybrary :**
     Une plateforme proposant des cours gratuits sur la cybersécurité et les domaines connexes. Visitez le site web de **Cybrary**.
   - **ArXiv :**
     Ressource inestimable pour les chercheurs, ArXiv est un dépôt d'articles préimprimés dans divers domaines, notamment l'IA et la cybersécurité. Consultez **ArXiv** pour vous tenir au courant des recherches en cours.
   - **Plateforme ANGELES du CCN-CERT**
     https://angeles.ccn-cert.cni.es/es/

7.3 Concevoir des systèmes robustes et résilients

La robustesse et la résilience dans la conception des systèmes, en particulier ceux qui intègrent l'intelligence artificielle, sont essentielles pour garantir qu'ils continuent à fonctionner de manière optimale dans des conditions défavorables et qu'ils peuvent se rétablir rapidement en cas de défaillance ou d'attaque.

Par **robustesse** nous entendons la capacité d'un système à résister aux perturbations et à continuer à fonctionner correctement sans dégradation. Un système robuste peut faire face à des conditions imprévues et à des variations dans l'environnement opérationnel.

D'un autre côté, la **résilience** est la capacité d'un système à se remettre rapidement d'une défaillance, à s'adapter à de nouvelles conditions et à rétablir son fonctionnement normal.

Pour obtenir des systèmes robustes et résilients, les **éléments suivants** doivent être pris en compte :

Principes de conception	**Réduction de la surface d'attaque :**	Réduire la surface d'attaque en limitant les points d'entrée dans le système et en éliminant les composants inutiles.
	Redondance :	Mise en œuvre de systèmes et de composants en double qui puissent faire face à la charge de travail en cas de défaillance d'un composant principal.
	Segmentation/ségrégation :	Division du système en segments plus petits et indépendants, de sorte qu'une défaillance ou une attaque dans un segment ne compromette pas l'ensemble du système.
	Contrôle/surveillance en continu :	Utiliser des outils et des solutions de contrôle et de surveillance pour détecter rapidement toute irrégularité ou anomalie.
	Mises à jour régulières :	Maintenir les logiciels et le matériel à jour afin de corriger les vulnérabilités connues.

Considérations relatives à l'IA	**Données d'entraînement robustes :**	Veillez à ce que les modèles d'IA soient entraînés à l'aide d'ensembles de données variés et actualisés afin de pouvoir gérer diverses situations.
	Validation rigoureuse :	Évaluer et valider les modèles dans différents scénarios et conditions.
	Défense contre les attaques adverses :	Mettre en œuvre des techniques telles que la régularisation et l'augmentation des données pour protéger les modèles d'IA contre les attaques qui cherchent à exploiter leurs faiblesses.
	Transparence et explicabilité :	Utiliser des modèles d'IA qui puissent être interprétés et vérifiés pour comprendre leur fonctionnement et leur prise de décision.
Tests et simulations	**Surveillance continue :**	Tests réguliers dans des environnements contrôlés afin d'identifier et de corriger les vulnérabilités.
	Simulations de défaillances :	Simuler des défaillances dans différentes parties du système afin d'évaluer le temps de réponse et de rétablissement.
	Exercices de réponse aux incidents :	S'entraîner, conformément à la réglementation applicable, à répondre à des incidents de sécurité potentiels afin d'améliorer l'efficience et l'efficacité dans des situations réelles.
Culture de l'amélioration continue	**Favoriser un état d'esprit consistant à chercher constamment à améliorer la robustesse et la résilience du système, en tirant les leçons des incidents et en s'adaptant aux nouvelles menaces et aux nouveaux défis.**	

En bref, la conception de systèmes robustes et résilients est essentielle pour garantir que les systèmes, en particulier ceux qui intègrent l'IA, puissent gérer des conditions défavorables et s'en remettre rapidement. Il s'agit d'une tâche permanente qui nécessite une combinaison de techniques de conception, de tests rigoureux et d'une culture d'amélioration constante.

8. Conclusion

8.1 Réflexions finales sur l'état actuel et l'avenir de l'IA dans la cybersécurité

L'évolution de la cybersécurité et de l'intelligence artificielle s'est révélée être un binôme fascinant, mais aussi un défi. Les deux disciplines, séparément, ont des trajectoires complexes, et leur intersection a provoqué à la fois des révolutions et des dilemmes. En réfléchissant à leur état actuel et au paysage futur, plusieurs **considérations clés** peuvent être tirées :

1. **Interdépendance croissante :**
   La cybersécurité ne peut plus être considérée comme une discipline indépendante de l'IA. L'immensité et la complexité du cyberespace, combinées à l'énorme quantité de données générées, font que les solutions basées sur l'IA soient essentielles pour une défense efficace.
2. **Des défis changeants :**
   Les menaces se multiplient à mesure que l'IA progresse. Les acteurs malveillants adoptent rapidement de nouvelles technologies pour améliorer leurs tactiques. C'est un jeu constant du chat et de la souris, où la défense et l'attaque évoluent en parallèle.

3. **Pertinence du facteur humain :**
   Malgré l'automatisation et les capacités avancées qu'apporte l'IA, le facteur humain reste irremplaçable. Les décisions éthiques, l'interprétation des données et la compréhension du contexte restent une responsabilité humaine. La collaboration homme-machine sera essentielle au succès de la cybersécurité à l'avenir.

4. **Défis éthiques et réglementaires :**
   L'adoption de l'IA dans le domaine de la cybersécurité s'accompagne de défis éthiques et réglementaires, comme dans le cas du règlement européen sur l'IA que nous avons évoqué à plusieurs reprises dans ce document. La vie privée (confidentialité), le consentement et la transparence sont des domaines qui doivent être abordés avec prudence et responsabilité, en particulier lorsqu'ils sont mis en balance avec le besoin de sécurité.

5. **Un potentiel inexploité :**
   Si l'IA appliquée à la cybersécurité a connu des avancées impressionnantes, il existe encore un vaste potentiel inexploité. Les technologies émergentes, telles que l'IA quantique et l'apprentissage fédéré, pourraient encore remodeler le paysage de la cybersécurité au cours de la prochaine décennie.

6. **Préparer l'avenir :**
   Les organisations et les professionnels de la cybersécurité doivent être prêts à s'adapter rapidement. La formation continue, la recherche et la collaboration interdisciplinaire seront essentielles pour se tenir au courant des dernières tendances et menaces.

7. **Vision holistique :**
   La cybersécurité, dans son essence, est une discipline holistique. Il ne s'agit plus seulement de technologie, mais aussi de processus, de personnes et de politiques. L'adoption de l'IA doit être considérée comme faisant partie d'une approche plus large et plus stratégique de la sécurisation du cyberespace.

En conclusion, l'imbrication de l'IA et de la cybersécurité redéfinit l'avenir de la sécurité numérique. Si elle offre des possibilités sans précédent pour une défense plus efficace et une détection plus rapide, elle introduit également des défis complexes qui doivent être relevés avec prudence, innovation et collaboration. La trajectoire future de cette intersection sera sans aucun doute passionnante et déterminante pour l'avenir numérique de l'humanité.

La collaboration homme-machine sera essentielle au succès de la cybersécurité à l'avenir

8.2 Actions ultérieures et recommandations pour la recherche future

L'évolution du paysage de la cybersécurité et de l'intelligence artificielle exige non seulement une réflexion sur ce que nous avons appris jusqu'à présent, mais aussi une vision claire des prochaines étapes. Alors que nous nous dirigeons vers un avenir plus numérique et interconnecté, il est essentiel que la communauté mondiale —des chercheurs aux professionnels, en passant par les législateurs et les citoyens ordinaires— s'unisse dans la mission de sécuriser notre cyberespace.

Certaines recommandations et actions subséquentes sont présentées ci-dessous :

1. **Création de centres de recherche collaboratifs :**
   Il est essentiel de créer davantage de centres et de plateformes permettant une collaboration interdisciplinaire dans le domaine de la cybersécurité et de l'IA. Ces centres peuvent servir de points de convergence pour la recherche innovante, en rassemblant des experts en IA, en cybersécurité, en droit et d'autres domaines connexes.
2. **Promouvoir l'éducation et la formation spécialisée :**
   Il est urgent d'élaborer des programmes éducatifs axés sur l'intersection de l'IA et de la cybersécurité. Il s'agit non seulement de remédier au manque de compétences dans ce domaine, mais aussi de veiller à ce que les futurs professionnels possèdent les connaissances nécessaires.
3. **Normes et standards mondiaux :**
   La communauté internationale devrait collaborer à l'élaboration de normes et de réglementations relatives à l'application de l'IA dans le domaine de la cybersécurité. Ces réglementations fourniront non seulement un cadre de référence, mais garantiront également que la technologie est utilisée de manière éthique et responsable. À cette fin, il est jugé essentiel de formaliser des cadres de certification pour les technologies, produits et services d'IA de confiance[44].

Alors que nous nous dirigeons vers un avenir plus numérique et interconnecté, il est essentiel que la communauté mondiale —des chercheurs aux professionnels, en passant par les législateurs et les citoyens ordinaires— s'unisse dans la mission de sécuriser notre cyberespace

44 Voir à ce sujet l'excellent travail de la Rand Corporation. *Labelling Initiatives, codes of conduct and others self-regulatory mechanisms for artificial intelligence applications* (2022).

4. **Recherche sur les menaces émergentes :**
   Avec l'évolution rapide de l'IA, les menaces changent et s'adaptent également. Il est essentiel de financer et de donner la priorité à la recherche sur les menaces émergentes, en particulier celles qui découlent des progrès technologiques les plus récents.

5. **Développement d'outils d'IA explicable :**
   Le monde a besoin de plus de recherche sur les outils et les méthodes qui rendent l'IA plus transparente et compréhensible. Les décisions prises par les algorithmes d'IA dans le domaine de la cybersécurité peuvent avoir des répercussions importantes, il est donc essentiel qu'elles puissent être expliquées et comprises.

6. **Promouvoir la vie privée et l'éthique :**
   Les recherches futures ne devraient pas seulement se concentrer sur l'efficacité et l'efficience des solutions d'IA dans le domaine de la cybersécurité, mais aussi sur leur impact en matière d'éthique et de respect de la vie privée. La protection de la vie privée ne doit pas être sacrifiée au nom de la sécurité.

7. **Essais et validations rigoureuses :**
   Avant de mettre en œuvre des solutions basées sur l'IA dans des environnements réels, il est essentiel de procéder à des essais et à des validations approfondies. Cela permettra de s'assurer que les solutions sont robustes et fiables face aux menaces du monde réel.

8. **Mesures d'incitation à l'innovation :**
   Les gouvernements et les organisations privées devraient fournir des mesures d'incitation à l'innovation en matière de cybersécurité et d'IA. Ces mesures peuvent prendre la forme de subventions, de concours ou de récompenses/reconnaissance.

En bref, nous nous trouvons à un point crucial à l'intersection de la cybersécurité et de l'IA. Alors que ces disciplines continuent d'évoluer et de s'entremêler, il est essentiel que nous adoptions une approche proactive, collaborative, réglementaire et éthique pour relever les défis de l'avenir. L'appel à l'action est clair : nous devons nous unir dans la mission de sécuriser notre avenir numérique, en protégeant nos données et nos infrastructures, et —en fin de compte— nos sociétés.

centro criptológico nacional

centro criptológico nacional

www.ccn.cni.es

www.ccn-cert.cni.es

https://oc.ccn.cni.es/


ccn-cert
centro criptológico nacional
CCN
centro criptológico nacional


PUBLICACIONES
Pd
DE DEFENSA


GOBIERNO
DE ESPAÑA
MINISTERIO
DE DEFENSA
SUBSECRETARÍA DE DEFENSA
SECRETARÍA GENERAL TÉCNICA
SUBDIRECCIÓN GENERAL
DE PUBLICACIONES
Y PATRIMONIO CULTURAL